

# Self-Healing Umbrella Sampling

## Convergence and efficiency

Gersende Fort · Benjamin Jourdain · Tony Lelièvre · Gabriel Stoltz

Received: date / Accepted: date

**Abstract** The Self-Healing Umbrella Sampling (SHUS) algorithm is an adaptive biasing algorithm which has been proposed in [16] in order to efficiently sample a multimodal probability measure. We show that this method can be seen as a variant of the well-known Wang-Landau algorithm [21,22]. Adapting results on the convergence of the Wang-Landau algorithm obtained in [8], we prove the convergence of the SHUS algorithm. We also compare the two methods in terms of efficiency. We finally propose a modification of the SHUS algorithm in order to increase its efficiency, and exhibit some similarities of SHUS with the well-tempered metadynamics method [2].

**Keywords** Wang-Landau algorithm · Stochastic Approximation Algorithm · Free energy biasing techniques

This work is supported by the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement number 614492 and by the French National Research Agency under the grant ANR-12-BS01-0019 (STAB). We also benefited from the scientific environment of the Laboratoire International Associé between the Centre National de la Recherche Scientifique and the University of Illinois at Urbana-Champaign.

G. Fort  
LTCI, CNRS & Telecom Paris Tech 46, rue Barrault  
75634 Paris Cedex 13  
France  
E-mail: gersende.fort@telecom-paristech.fr

B. Jourdain, T. Lelièvre and G. Stoltz  
Université Paris-Est, CERMICS (ENPC), INRIA  
F-77455 Marne-la-Vallée  
France E-mail: {jourdain,lelievre,stoltz}@cermics.enpc.fr

## 1 Introduction

The efficient sampling of a probability measure defined over a high dimensional space is required in many application fields, such as computational statistics or molecular dynamics [15]. Standard algorithms consist in building a dynamics which is ergodic with respect to the target distribution such as Langevin dynamics [15,20] or Metropolis-Hastings dynamics [17,12]. Averages over trajectories of this ergodic dynamics are then used as approximations of averages with respect to the target probability measure. In many cases of interest, this probability measure is multimodal: regions of high probability are separated by regions of low probability, and this implies that the ergodic dynamics is metastable. This means that it takes a lot of time to leave a high probability region (called a metastable state). The consequence of this metastable behavior is that trajectorial averages converge very slowly to their ergodic limits.

Many techniques have been proposed to overcome these difficulties. Among them, importance sampling consists in modifying the target probability using a well-chosen bias in order to enhance the convergence to equilibrium. Averages with respect to the original target are then recovered using a reweighting of the biased samples. In general, it is not easy to devise an appropriate bias. Adaptive importance sampling methods have thus been proposed in order to automatically build a “good” bias (see [15, Chapter 5] for a review of these approaches).

Let us explain the principle of adaptive biasing techniques in a specific setting (we refer to [15, Chapter 5] for a more general introduction to such methods). To fix the ideas, let us consider a target probability measure  $\pi d\lambda$  on the state space  $X \subseteq \mathbb{R}^D$ , where  $\lambda$  denotes the Lebesgue measure on  $\mathbb{R}^D$ , and let us be given a

partition  $X_1, \dots, X_d$  of  $X$  into  $d$  subsets. The subsets  $X_i$  are henceforth called 'strata'. The choice of such a partition will be discussed later (see Footnote 1 below). We assume that the target measure is multimodal in the sense that the weights of the strata span several orders of magnitude. In other words, the weights of the strata

$$\theta_*(i) = \int_{X_i} \pi(x) d\lambda(x), \quad i = 1, \dots, d \quad (1)$$

vary a lot and  $\frac{\max_{1 \leq i \leq d} \theta_*(i)}{\min_{1 \leq i \leq d} \theta_*(i)}$  is large. We suppose from now on and without loss of generality (up to removing some strata) that  $\min_{1 \leq i \leq d} \theta_*(i) > 0$ . In such a situation, it is natural to consider the following biased probability density:

$$\pi_{\theta_*}(x) = \frac{1}{d} \sum_{i=1}^d \frac{\pi(x)}{\theta_*(i)} \mathbf{1}_{X_i}(x), \quad (2)$$

which is such that

$$\int_{X_i} \pi_{\theta_*}(x) d\lambda(x) = \frac{1}{d} \quad (3)$$

for all  $i \in \{1, \dots, d\}$ : under the biased probability measure  $\pi_{\theta_*} d\lambda$ , each stratum has the same weight. In particular, the ergodic dynamics which are built with  $\pi_{\theta_*}$  as the target measure are typically less metastable than the dynamics with target  $\pi$ . The practical difficulty to implement this technique is of course that the vector

$$\theta_* = \{\theta_*(1), \dots, \theta_*(d)\}$$

is unknown. The principle of adaptive biasing methods is to learn on the fly the vector  $\theta_*$  in order to eventually sample the biased probability measure  $\pi_{\theta_*}$ . Adaptive algorithms thus build a sequence of vectors  $(\theta_n)_{n \geq 0}$  which is expected to converge to  $\theta_*$ . Various updating procedures have been proposed [15, Chapter 5]. Such adaptive techniques are used on a daily basis by many practitioners in particular for free energy computations in computational statistical physics. In this context, the partition of  $X$  is related to the choice of a so-called reaction coordinate<sup>1</sup>, and the weights  $\theta_*$  give the free energy profile associated with this reaction coordinate. We focus here on a specific adaptive biasing method called the Self-Healing Umbrella Sampling (SHUS) technique [16,6]. We will show that it is a variant of the well-known Wang-Landau method [21,22].

<sup>1</sup> For a given measure  $\pi d\lambda$ , the choice of a good partition or, equivalently, of a good reaction coordinate does of course influence the efficiency of the algorithm. This choice is a difficult problem that we do not consider here (see for example [4] for such discussions in the context of computational statistics). Here, both  $\pi d\lambda$  and  $X_1, \dots, X_d$  are assumed to be given.

From a practical viewpoint, the main interest of SHUS compared to Wang-Landau is that the practitioner has less numerical parameters to tune (as will be explained in Section 2.3).

The aim of this paper is to analyze the SHUS algorithm in terms of convergence and efficiency. First, we adapt the results of [8] which prove the convergence of Wang-Landau to obtain the convergence of SHUS (see Theorem 2). Second, we perform numerical experiments to analyze the efficiency of SHUS, in the spirit of [9] where similar numerical tests are performed for the Wang-Landau algorithm. The efficiency analysis consists in estimating the average exit time from a metastable state in the limit when  $\frac{\max_{1 \leq i \leq d} \theta_*(i)}{\min_{1 \leq i \leq d} \theta_*(i)}$  goes to infinity. Adaptive techniques (such as SHUS or Wang-Landau) yield exit times which are much smaller than for the original non-adaptive dynamics.

The main output of this work is that, both in terms of convergence (longtime behavior) and efficiency (exit times from metastable states), SHUS is essentially equivalent to the Wang-Landau algorithm for a specific choice of the numerical parameters. These numerical parameters are not the optimal ones in terms of efficiency and we propose in Section 5.2 a modified SHUS algorithm which is (in the longtime limit) equivalent to the Wang-Landau algorithm with better sets of parameters.

This article is organized as follows. In Section 2, we introduce the SHUS algorithm, check its asymptotic correctness and explain how it can be seen as a Wang-Landau algorithm with stochastic stepsize sequence. In Section 3, we state a convergence result for Wang-Landau algorithms with general (either deterministic or stochastic) stepsize sequences and deduce the convergence of SHUS. The proofs are based on stochastic approximation arguments and postponed to Section 6. Numerical results illustrating the efficiency of the algorithm and comparing its performance with the standard Wang-Landau algorithm are provided in Section 4. Finally, in Section 5, we draw some conclusions on the interest of SHUS compared with the Wang-Landau algorithm, and further compare SHUS with other adaptive techniques, such as the well-tempered metadynamics algorithm [2] and the above-mentioned modified SHUS algorithm. We also prove the convergence of this modified SHUS algorithm (see Proposition 4), and present numerical results showing that this new method is closely related to a Wang-Landau dynamics with larger stepsizes.

## 2 The SHUS algorithm

Using the notation of the introduction, we consider a target probability measure  $\pi d\lambda$  on the state space  $\mathbf{X} \subseteq \mathbb{R}^D$  and a partition of  $\mathbf{X}$  into  $d$  strata  $\mathbf{X}_1, \dots, \mathbf{X}_d$ .

We introduce a family of biased densities  $\pi_\theta$ , for  $\theta \in \Theta$ , where  $\Theta$  is the set of positive probability measures on  $\{1, \dots, d\}$ :

$$\Theta = \left\{ \theta = (\theta(1), \dots, \theta(d)) \in (0, 1)^d, \sum_{i=1}^d \theta(i) = 1 \right\}.$$

The biased densities  $(\pi_\theta)_{\theta \in \Theta}$  are obtained from  $\pi$  by a reweighting of each stratum:

$$\pi_\theta(x) = \left( \sum_{j=1}^d \frac{\theta_\star(j)}{\theta(j)} \right)^{-1} \sum_{i=1}^d \frac{\pi(x)}{\theta(i)} \mathbf{1}_{\mathbf{X}_i}(x), \quad (4)$$

where  $\theta_\star \in \Theta$  is defined by (1). Observe that for any  $\theta \in \Theta$  and  $i \in \{1, \dots, d\}$ ,

$$\int_{\mathbf{X}_i} \pi_\theta(x) d\lambda(x) = \frac{\theta_\star(i)/\theta(i)}{\sum_{j=1}^d \theta_\star(j)/\theta(j)}. \quad (5)$$

Equations (2) and (3) are respectively Equations (4) and (5) with the specific choice  $\theta = \theta_\star$ .

### 2.1 Description of the algorithm

As explained above, the principle of many adaptive biasing techniques, and SHUS in particular, is to build a sequence  $(\theta_n)_{n \geq 1}$  which converges to  $\theta_\star$ . This allows to sample  $\pi_{\theta_\star}$ , which is less multimodal than  $\pi$ . In order to understand how the updating rule for  $\theta_n$  is built for SHUS, one may proceed as follows.

Let us first assume that we are given a Markov chain  $(X_n)_{n \geq 0}$  which is ergodic with respect to the target measure  $\pi d\lambda$  (think of a Metropolis-Hastings dynamics). Let us introduce the sequence (for a given  $\gamma > 0$ )

$$\begin{aligned} \tilde{\theta}_{n+1}(i) &= \tilde{\theta}_n(i) + \gamma \mathbf{1}_{\mathbf{X}_i}(X_{n+1}) \\ &= \begin{cases} \tilde{\theta}_n(i) + \gamma & \text{if } X_{n+1} \in \mathbf{X}_i, \\ \tilde{\theta}_n(i) & \text{otherwise,} \end{cases} \end{aligned} \quad (6)$$

which, in some sense, counts the number of visits to each stratum. By the ergodic property, it is straightforward to check that  $\theta_n = \frac{\tilde{\theta}_n}{\sum_{j=1}^d \tilde{\theta}_n(j)}$  converges almost surely (a.s.) to  $\theta_\star$  as  $n \rightarrow \infty$ . As explained in the introduction, the difficulty with this algorithm is that the convergence of  $\theta_n$  to  $\theta_\star$  is very slow due to the metastability of the density  $\pi$  and thus of the Markov chain  $(X_n)_{n \geq 0}$ .

The idea is then that if an estimate  $\bar{\theta}$  of  $\theta_\star$  is available, one should instead consider a Markov chain  $(X_n)_{n \geq 0}$  which is ergodic with respect to  $\pi_{\bar{\theta}}$  and thus hopefully less metastable. To estimate  $\theta_\star$  with this new Markov chain, one should modify the updating rule (6) as

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) + \gamma \bar{\theta}(i) \mathbf{1}_{\mathbf{X}_i}(X_{n+1}) \quad (7)$$

in order to unbiased the samples  $(X_n)_{n \geq 0}$  (since  $\frac{\pi(x)}{\pi_{\bar{\theta}}(x)} = \left( \sum_{j=1}^d \frac{\theta_\star(j)}{\bar{\theta}(j)} \right) \sum_{i=1}^d \bar{\theta}(i) \mathbf{1}_{\mathbf{X}_i}(x)$ ). Again, by the ergodic property, one easily gets that  $\theta_n = \frac{\tilde{\theta}_n}{\sum_{j=1}^d \tilde{\theta}_n(j)}$  converges a.s. to  $\theta_\star$  as  $n \rightarrow \infty$ . Indeed,  $n^{-1} \sum_{k=1}^n \mathbf{1}_{\mathbf{X}_i}(X_k)$  converges a.s. to  $\int_{\mathbf{X}_i} \pi_{\bar{\theta}}(x) d\lambda(x)$  (given by (5)) as  $n \rightarrow \infty$ . Since

$$\frac{\tilde{\theta}_n(i)}{n} = \frac{\tilde{\theta}_0(i)}{n} + \gamma \frac{\bar{\theta}(i)}{n} \sum_{k=1}^n \mathbf{1}_{\mathbf{X}_i}(X_k), \quad (8)$$

$(\tilde{\theta}_n(i)/n)_{n \geq 0}$  converges a.s. to  $\frac{\gamma \theta_\star(i)}{\sum_{j=1}^d \theta_\star(j)/\bar{\theta}(j)}$ , which implies in turn

$$\lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^d \tilde{\theta}_n(i) = \frac{\gamma}{\sum_{j=1}^d \theta_\star(j)/\bar{\theta}(j)} \quad \text{a.s.}$$

Hence  $\theta_n = \frac{\tilde{\theta}_n}{\sum_{j=1}^d \tilde{\theta}_n(j)}$  converges a.s. to  $\theta_\star$  as  $n \rightarrow \infty$ .

The SHUS algorithm consists in using the current value  $\theta_n$  as the estimate  $\bar{\theta}$  of  $\theta_\star$  in the previous algorithm. Let us now precisely define the SHUS algorithm. Let  $(P_\theta)_{\theta \in \Theta}$  be a family of transition kernels on  $\mathbf{X}$  which are ergodic with respect to  $\pi_\theta$ . In particular,

$$\forall \theta \in \Theta, \quad \pi_\theta P_\theta = \pi_\theta.$$

Let  $\gamma > 0$ ,  $X_0 \in \mathbf{X}$  and  $\tilde{\theta}_0 = (\tilde{\theta}_0(1), \dots, \tilde{\theta}_0(d)) \in (\mathbb{R}_+^*)^d$  be deterministic. The SHUS algorithm consists in iterating the following steps:

**Algorithm 1** Given  $(\tilde{\theta}_n, X_n) \in (\mathbb{R}_+^*)^d \times \mathbf{X}$ ,

- compute the probability measure on  $\{1, \dots, d\}$ ,

$$\theta_n = \frac{\tilde{\theta}_n}{\sum_{j=1}^d \tilde{\theta}_n(j)} \in \Theta, \quad (9)$$

- draw  $X_{n+1}$  according to the kernel  $P_{\theta_n}(X_n, \cdot)$ ,
- compute, for all  $i \in \{1, \dots, d\}$ ,

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) + \gamma \theta_n(i) \mathbf{1}_{\mathbf{X}_i}(X_{n+1}). \quad (10)$$

Notice that only simulation under  $P_\theta$  is needed to implement SHUS. By choosing  $P_\theta$  as a Metropolis-Hastings kernel with target measure  $\pi_\theta d\lambda$ , the density  $\pi_\theta$  needs to be known only up to a normalizing constant. Therefore, SHUS covers the case when the target measure  $\pi d\lambda$  is only known up to a multiplicative constant which

is generally the case in view of applications to Bayesian statistics and statistical physics.

As proved in Section 3,  $(\theta_n)_{n \geq 0}$  converges almost surely to  $\theta_*$  when  $n \rightarrow \infty$ . According to the update mechanism (10),  $\tilde{\theta}_{n+1}(i) - \tilde{\theta}_n(i)$  is non negative, and it is positive if and only if the current draw  $X_{n+1}$  lies in stratum  $i$ . In addition, this variation is proportional to the current approximation  $\theta_n(i)$  of  $\theta_*(i)$  with a factor  $\gamma$  chosen by the user (prior to the run of the algorithm; the choice of  $\gamma$  is numerically investigated in Section 4).

The principle of this algorithm is thus to penalize the already visited strata, in order to favor transitions towards unexplored strata. The penalization strength is proportional to the current bias of the strata. The prefactor  $\theta_n(i)$  in (10) can be understood as a way to unbiased the samples  $(X_n)_{n \geq 0}$  in order to recover samples distributed according to the target measure  $\pi \, d\lambda$ , see also formula (17) below.

## 2.2 Asymptotic correctness

Let us consider the SHUS algorithm 1 and let us assume that  $(\theta_n)_{n \geq 0}$  converges to some value, say  $\bar{\theta} \in \Theta$ . It is then expected that  $n^{-1} \sum_{k=1}^n \mathbf{1}_{X_k}(X_k)$  converges to  $\int_{X_i} \pi_{\bar{\theta}}(x) \, d\lambda(x)$  and that the updating rule (10) leads to the same asymptotic behavior as (7) (where  $\theta_n(i)$  in (10) has been replaced by its limit  $\bar{\theta}(i)$ ). Thus, it is expected that  $\lim_n n^{-1} \tilde{\theta}_n = \frac{\gamma \theta_*}{\sum_{j=1}^d \theta_*(j)/\bar{\theta}(j)}$  a.s. by (8) and thus  $\lim_n \theta_n = \theta_*$  a.s.. Therefore, the only possible limit of the sequence  $(\theta_n)_{n \geq 0}$  is  $\theta_*$ .

This heuristic argument is of course not a proof of convergence, but it explains why one can expect the SHUS algorithm to behave like a Metropolis-Hastings algorithm with target measure  $\pi_{\theta_*}$ . The rigorous result for the convergence is given in Section 3, and the efficiency of SHUS is discussed in Section 4.

## 2.3 Reformulation as a Wang-Landau algorithm with a stochastic stepsize sequence

One key observation of this work is that SHUS can be seen as a Wang-Landau algorithm with nonlinear update of the weights and with a specific stepsize sequence  $(\gamma_n)_{n \geq 1}$ , see [21, 22, 8, 9]. The Wang-Landau algorithm with nonlinear update of the weights consists in replacing the updating formula (10) by:

$$\tilde{\theta}_{n+1}^{\text{WL}}(i) = \tilde{\theta}_n^{\text{WL}}(i) \left( 1 + \gamma_{n+1}^{\text{WL}} \mathbf{1}_{X_i}(X_{n+1}) \right), \quad (11)$$

where the deterministic stepsize sequence  $(\gamma_n^{\text{WL}})_{n \geq 1}$  has to be *chosen* by the practitioner beforehand. The choice of this sequence is not easy: it should converge to zero

when  $n$  goes to infinity (vanishing adaption) in order to ensure the convergence of the sequence  $(\tilde{\theta}_n^{\text{WL}})_{n \geq 0}$ , but not too fast otherwise the convergence of  $\theta_n^{\text{WL}} = \frac{\tilde{\theta}_n^{\text{WL}}}{\sum_{i=1}^d \tilde{\theta}_n^{\text{WL}}(i)}$  to  $\theta_*$  is not ensured.

*Remark 1* In fact, the updating rule of the original Wang-Landau algorithm is more complicated than (11) since the stepsizes are changed at random stopping times related to a quasi-uniform population of the strata and not at every iteration, see [13] for a mathematical analysis of the well-posedness of the algorithm.

Going back to SHUS, by setting

$$\gamma_{n+1} = \frac{\gamma}{\sum_{j=1}^d \tilde{\theta}_n(j)}, \quad \text{for } n \in \mathbb{N}, \quad (12)$$

it is easy to check that (10) is equivalent to

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) \left( 1 + \gamma_{n+1} \mathbf{1}_{X_i}(X_{n+1}) \right), \quad (13)$$

which explains why SHUS can be seen as a Wang-Landau algorithm with nonlinear update of the weights (see (11)) but with a stepsize sequence  $(\gamma_n)_{n \geq 1}$  which is not chosen by the practitioner: it is adaptively built by the algorithm.

## 3 Convergence results

The goal of this section is to establish convergence results on the sequence  $(\theta_n)_{n \geq 0}$  given by (9) to the weight vector  $\theta_*$  and on the distribution of the samples  $(X_n)_{n \geq 0}$ . To do so, we will extend the results of [8] and more generally prove convergence of the Wang-Landau algorithm with general (either deterministic or stochastic) stepsize sequence  $(\gamma_n)_{n \geq 1}$  and the nonlinear update of the weights. We need the following assumptions on the target density  $\pi$  and the kernels  $P_\theta$ :

**A1** The density  $\pi$  of the target distribution is such that  $0 < \inf_X \pi \leq \sup_X \pi < \infty$  and the strata  $(X_i)_{i \in \{1, \dots, d\}}$  satisfy  $\min_{1 \leq i \leq d} \lambda(X_i) > 0$ .

Notice that this assumption implies that  $\theta_*$  given by (1) is such that  $\min_{1 \leq i \leq d} \theta_*(i) > 0$  (hence  $\theta_* \in \Theta$ ).

**A2** For any  $\theta \in \Theta$ ,  $P_\theta$  is a Metropolis-Hastings transition kernel with proposal kernel  $q(x, y) \, d\lambda(y)$  where  $q(x, y)$  is symmetric and satisfies  $\inf_{x \in X} q > 0$ , and with invariant distribution  $\pi_\theta \, d\lambda$ , where  $\pi_\theta$  is given by (4).

### 3.1 Convergence of Wang-Landau algorithms with a general stepsize sequence

In [8], we consider the Wang-Landau algorithm with a linear update of the weights. The linear update version of Wang-Landau consists in changing the updating rule (11) to a linearized version (in the limit  $\gamma_{n+1}^{\text{WL}} \rightarrow 0$ ) on the normalized weights:

$$\theta_{n+1}^{\text{WL}}(i) = \theta_n^{\text{WL}}(i) + \gamma_{n+1}^{\text{WL}} \theta_n^{\text{WL}}(i) (\mathbf{1}_{X_i}(X_{n+1}) - \theta_n^{\text{WL}}(I(X_{n+1}))). \quad (14)$$

We prove in [8] that, when the target density  $\pi$  and the kernels  $P_\theta$  satisfy A1 and A2, the Wang-Landau algorithm with this linear update of the weights converges under the following condition on  $(\gamma_n^{\text{WL}})_{n \geq 1}$ : the sequence  $(\gamma_n^{\text{WL}})_{n \geq 1}$  is deterministic, ultimately non-increasing,

$$\sum_{n \geq 1} \gamma_n^{\text{WL}} = +\infty \text{ and } \sum_{n \geq 1} (\gamma_n^{\text{WL}})^2 < +\infty.$$

To the best of knowledge, neither the convergence of the Wang-Landau algorithm with the nonlinear update of the weights (11) nor the case of a random sequence of stepsizes are addressed in the literature. It is a particular case of the following Wang-Landau algorithm with general stepsize sequence which also generalizes SHUS: starting from random variables  $\tilde{\theta}_0 = (\tilde{\theta}_0(1), \dots, \tilde{\theta}_0(d)) \in (\mathbb{R}_+^*)^d$  and  $X_0 \in \mathbf{X}$ , iterate the following steps:

**Algorithm 2** Given  $(\tilde{\theta}_n, X_n) \in (\mathbb{R}_+^*)^d \times \mathbf{X}$ ,

– compute the probability measure on  $\{1, \dots, d\}$ ,

$$\theta_n = \frac{\tilde{\theta}_n}{\sum_{j=1}^d \tilde{\theta}_n(j)} \in \Theta, \quad (15)$$

– draw  $X_{n+1}$  according to the kernel  $P_{\theta_n}(X_n, \cdot)$ ,  
– compute, for all  $i \in \{1, \dots, d\}$ ,

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i)(1 + \gamma_{n+1} \mathbf{1}_{X_i}(X_{n+1})), \quad (16)$$

where the positive stepsize sequence  $(\gamma_n)_{n \geq 1}$  is supposed to be predictable with respect to the filtration  $\mathcal{F}_n = \sigma(\tilde{\theta}_0, X_0, X_1, \dots, X_n)$  (i.e.  $\gamma_n$  is  $\mathcal{F}_{n-1}$ -measurable).

Algorithm 2 is a meta-algorithm: to obtain a practical algorithm, one has to specify the way the stepsize sequence  $(\gamma_n)_{n \geq 1}$  is generated.

**Definition 1** An algorithm of the type described in Algorithm 2 is said to converge if it satisfies the following properties:

$$(i) \quad \mathbb{P} \left( \lim_{n \rightarrow +\infty} \theta_n = \theta_\star \right) = 1.$$

(ii) For any bounded measurable function  $f$  on  $\mathbf{X}$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}[f(X_n)] &= \int_{\mathbf{X}} f(x) \pi_{\theta_\star}(x) d\lambda(x), \\ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(X_k) &= \int_{\mathbf{X}} f(x) \pi_{\theta_\star}(x) d\lambda(x) \quad \text{a.s.} \end{aligned}$$

(iii) For any bounded measurable function  $f$  on  $\mathbf{X}$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} \left[ d \sum_{j=1}^d \theta_{n-1}(j) f(X_n) \mathbf{1}_{X_j}(X_n) \right] &= \int_{\mathbf{X}} f(x) \pi(x) d\lambda(x), \\ \lim_{n \rightarrow \infty} \frac{d}{n} \sum_{k=1}^n \sum_{j=1}^d \theta_{k-1}(j) \mathbf{1}_{X_j}(X_k) f(X_k) &= \int_{\mathbf{X}} f(x) \pi(x) d\lambda(x) \quad \text{a.s.} \end{aligned} \quad (17)$$

In addition to the almost sure convergence of  $(\theta_n)_{n \geq 0}$  to  $\theta_\star$ , this definition encompasses the ergodic convergence of the random sequence  $(X_n)_{n \geq 0}$  to  $\pi_{\theta_\star} d\lambda$  and the convergence of an importance sampling-type Monte Carlo average to the probability measure  $\pi d\lambda$ . Our main result is the following Theorem.

**Theorem 1** Assume A1, A2 and that there exists a deterministic sequence  $(\bar{\gamma}_n)_n$  converging to 0 such that

$$\mathbb{P} \left( \forall n \geq 1, \gamma_n \leq \bar{\gamma}_n, \sum_n \gamma_n = +\infty, \sum_n \gamma_n^2 < \infty \right) = 1, \quad (18)$$

and the stepsize sequence  $(\gamma_n)_{n \geq 1}$  is a.s. non-increasing. Then Algorithm 2 converges in the sense of Definition 1.

One immediately deduces convergence of the Wang-Landau algorithm with deterministic stepsize sequence and nonlinear update (11) under the same assumptions as for the linear update (14), which generalizes the result of [8].

**Corollary 1** Assume A1, A2 and that the deterministic sequence  $(\gamma_n^{\text{WL}})_{n \geq 1}$  is non-increasing and such that  $\sum_{n \geq 1} \gamma_n^{\text{WL}} = +\infty$  and  $\sum_{n \geq 1} (\gamma_n^{\text{WL}})^2 < \infty$ . Then the Wang-Landau algorithm with nonlinear update (11) converges in the sense of Definition 1.

Before outlining the proof of Theorem 1, let us discuss its application to the SHUS algorithm.

### 3.2 Convergence of SHUS

The stepsize sequence  $(\gamma_n)_{n \geq 1} = \left( \frac{1}{\sum_{i=1}^d \tilde{\theta}_{n-1}(i)} \right)_{n \geq 1}$  obtained in the reformulation of the SHUS algorithm 1 as a Wang-Landau algorithm is clearly decreasing since, the sequence  $(\sum_{i=1}^d \tilde{\theta}_n(i))_{n \geq 0}$  is increasing. To apply Theorem 1, we also need to check (18). This is the purpose of the following proposition which is proved in Section 6.

**Proposition 1** *With probability one, the stepsize sequence  $(\gamma_n)_{n \geq 1}$  in the SHUS algorithm 1 is decreasing and for any  $n \in \mathbb{N}$ ,*

$$\forall n \in \mathbb{N}, \quad \frac{\gamma_1}{1 + n\gamma_1} \leq \gamma_{n+1} \leq \frac{\gamma_1}{\sqrt{1 + 2n\gamma_1 \min_{1 \leq i \leq d} \theta_0(i)}}.$$

Moreover, under A1 and A2, there exists a random variable  $C$  such that

$$\mathbb{P} \left( C > 0 \text{ and } \sup_{n \in \mathbb{N}} n^{\frac{1+C}{2}} \gamma_{n+1} < \infty \right) = 1.$$

Since  $\gamma_1 = \frac{\gamma}{\sum_{i=1}^d \tilde{\theta}_0(i)}$  where  $\tilde{\theta}_0$  is deterministic, combining Theorem 1 and Proposition 1, we obtain the following convergence result for the SHUS algorithm 1.

**Theorem 2** *Under A1 and A2, the SHUS algorithm 1 converges in the sense of Definition 1.*

*Remark 2* Since the sequence  $((X_n, \tilde{\theta}_n))_{n \geq 0}$  generated by the SHUS algorithm is a Markov chain, one easily deduces that the SHUS algorithm started from a random initial condition  $(X_0, \tilde{\theta}_0) \in \mathbf{X} \times (\mathbb{R}_+^*)^d$  also converges in the sense of Definition 1.

The convergence result from Theorem 2 allows us to characterize the asymptotic behavior of the stepsizes. Indeed, since for  $i \in \{1, \dots, d\}$  and  $n \geq 1$ ,

$$\frac{\tilde{\theta}_n(i)}{n} = \frac{\tilde{\theta}_0(i)}{n} + \frac{\gamma}{n} \sum_{k=1}^n \theta_{k-1}(i) \mathbf{1}_{\mathbf{X}_i}(X_k),$$

the property (17) with  $f(x) = \mathbf{1}_{\mathbf{X}_i}(x)$  implies that the SHUS algorithm generates sequences  $(\tilde{\theta}_n)_{n \geq 0}$  and  $(\gamma_n)_{n \geq 1}$  which satisfy

**Corollary 2** *Under A1 and A2,*

$$\lim_{n \rightarrow \infty} \frac{\tilde{\theta}_n}{n} = \frac{\gamma \theta_*}{d} \quad \text{a.s.} \quad \text{and} \quad \lim_{n \rightarrow \infty} n\gamma_n = d \quad \text{a.s.}$$

Notice that this Corollary implies that the stepsize sequence  $(\gamma_n)_{n \geq 1}$  scales like  $d/n$  in the large  $n$  limit. In Section 4.3, we will therefore compare SHUS with the Wang-Landau algorithm implemented with a stepsize sequence  $\gamma_n^{\text{WL}} = \frac{\gamma_*}{n}$ , for some positive parameter  $\gamma_*$ .

### 3.3 Strategy of the proof of Theorem 1

The proof of Theorem 1 is given in Section 6. It relies on a rewriting of the updating mechanism of the sequence  $(\theta_n)_{n \geq 0}$  as a Stochastic Approximation (SA) algorithm for which convergence results have been proven (see for example [1]). Notice that the updating formula (15)–(16) is equivalent to: for all  $i \in \{1, \dots, d\}$  and  $n \in \mathbb{N}$

$$\theta_{n+1}(i) = \theta_n(i) \frac{1 + \gamma_{n+1} \mathbf{1}_{\mathbf{X}_i}(X_{n+1})}{1 + \gamma_{n+1} \theta_n(I(X_{n+1}))}, \quad (19)$$

where for all  $x \in \mathbf{X}$ ,

$$I(x) = \sum_{j=1}^d j \mathbf{1}_{\mathbf{X}_j}(x)$$

denotes the index of the stratum where  $x$  lies. Upon noting that  $(1+a)/(1+b) = 1 + a - b + b(b-a)/(1+b)$ , (19) is equivalent to

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} H_i(X_{n+1}, \theta_n) + \gamma_{n+1} \Lambda_{n+1}(i), \quad (20)$$

where  $H : \mathbf{X} \times \Theta \rightarrow \mathbb{R}^d$  is defined by

$$H_i(x, \theta) = \theta(i) \left( \mathbf{1}_{\mathbf{X}_i}(x) - \theta(I(x)) \right), \quad (21)$$

and

$$\begin{aligned} \Lambda_{n+1}(i) &= \gamma_{n+1} \theta_n(i) \theta_n(I(X_{n+1})) \\ &\quad \times \frac{\theta_n(I(X_{n+1})) - \mathbf{1}_{\mathbf{X}_i}(X_{n+1})}{1 + \gamma_{n+1} \theta_n(I(X_{n+1}))}. \end{aligned} \quad (22)$$

Notice that the last term  $\gamma_{n+1} \Lambda_{n+1}$  in (20) is of the order of  $\gamma_{n+1}^2$ . The recurrence relation (20) is thus in a standard form to apply convergence results for SA algorithms (see *e.g.* [1]).

There are however three specific points in Algorithm 2 which make the study of this SA recursion quite technical. A first difficulty raises from the fact that  $(X_n)_{n \geq 0}$  alone is not a Markov chain: given the past up to time  $n$ ,  $X_{n+1}$  is generated according to a Markov transition kernel computed at the current position  $X_n$  but controlled by the current value  $\theta_n$ , which depends on the whole trajectory  $(\tilde{\theta}_0, X_0, \dots, X_n)$ . A second one comes from the randomness of the stepsizes  $(\gamma_n)_{n \geq 1}$ . Finally, it is not clear whether the sequence  $(\theta_n)_{n \geq 0}$  remains a.s. in a compact subset of the open set  $\Theta$  so that a preliminary step when proving the convergence of the SA recursion is to establish its recurrence, namely the fact that the sequence  $(\theta_n)_{n \geq 0}$  returns to a compact set of  $\Theta$  infinitely often.

### 3.4 A few crucial intermediate results

Let us highlight a few results which are crucial to tackle these difficulties and establish Theorem 1. The fundamental result to address the dynamics of  $(X_n)_{n \geq 0}$  controlled by  $(\theta_n)_{n \geq 0}$  is the following proposition established in [8, Proposition 3.1.]

**Proposition 2** *Assume A1 and A2. There exists  $\rho \in (0, 1)$  such that*

$$\sup_{x \in \mathbf{X}} \sup_{\theta \in \Theta} \|P_\theta^n(x, \cdot) - \pi_\theta d\lambda\|_{\text{TV}} \leq 2(1 - \rho)^n,$$

where for a signed measure  $\mu$ , the total variation norm is defined as

$$\|\mu\|_{\text{TV}} = \sup_{\{f: \sup_{\mathbf{X}} |f| \leq 1\}} |\mu(f)|.$$

In the present case, the recurrence property of the SA algorithm means the existence of a positive threshold such that infinitely often in  $n$ , the minimal weight

$$\underline{\theta}_n = \min_{1 \leq i \leq d} \theta_n(i) \quad (23)$$

is larger than the threshold. Let

$$I_n = \min \{i : \theta_n(i) = \underline{\theta}_n\}, \quad (24)$$

be the smallest index of stratum with smallest weight according to  $\theta_n$  and  $(T_k)_{k \in \mathbb{N}}$  be the times of return to the stratum of smallest weight:  $T_0 = 0$  and, for  $k \geq 1$ ,

$$T_k = \inf \{n > T_{k-1} : X_n \in \mathbf{X}_{I_n}\}, \quad (25)$$

with the convention  $\inf \emptyset = +\infty$ . We prove in Section 6 the following recurrence property.

**Proposition 3** *Assume A1, A2 and the existence of a deterministic sequence  $(\bar{\gamma}_n)_n$  converging to 0 such that  $\mathbb{P}(\forall n \geq 1, \gamma_n \leq \bar{\gamma}_n) = 1$ . Then Algorithm 2 is such that*

$$\mathbb{P}\left(\forall k \in \mathbb{N}, T_k < +\infty \text{ and } \limsup_{k \rightarrow \infty} \underline{\theta}_{T_k-1} > 0\right) = 1, \quad (26)$$

and

$$\mathbb{P}\left(\exists C_T < +\infty, \forall k \in \mathbb{N}, T_k \leq C_T k\right) = 1. \quad (27)$$

Using the recurrence property of Proposition 3, we are then able to prove that Algorithm 2 converges in the sense of Definition 1-(i) by using general convergence results for SA algorithms given in [1]. The properties in Definition 1-(ii)-(iii) then follow from convergence results for adaptive Markov Chain Monte Carlo algorithms given in [10]. See Section 6 for the details.

## 4 Numerical investigation of the efficiency

We present in this section some numerical results illustrating the efficiency of SHUS in terms of exit times from a metastable state. We also compare the performances of SHUS and of the Wang-Landau algorithm on this specific example.

### 4.1 Presentation of the model and of the dynamics

We consider the system based on the two-dimensional potential suggested in [18], see also [9] for similar experiments on the Wang-Landau algorithm. The state space is  $\mathbf{X} = [-R, R] \times \mathbb{R}$  (with  $R = 1.2$ ), and we denote by  $x = (x_1, x_2)$  a generic element of  $\mathbf{X}$ . The reference measure  $\lambda$  is the Lebesgue measure  $dx_1 dx_2$ . The density of the target measure reads

$$\pi(x) = Z^{-1} \mathbf{1}_{[-R, R]}(x_1) e^{-\beta U(x_1, x_2)},$$

for some positive inverse temperature  $\beta$ , with

$$\begin{aligned} U(x_1, x_2) = & 3 \exp\left(-x_1^2 - \left(x_2 - \frac{1}{3}\right)^2\right) - 3 \exp\left(-x_1^2 - \left(x_2 - \frac{5}{3}\right)^2\right) \\ & - 5 \exp\left(-(x_1 - 1)^2 - x_2^2\right) - 5 \exp\left(-(x_1 + 1)^2 - x_2^2\right) \\ & + 0.2x_1^4 + 0.2\left(x_2 - \frac{1}{3}\right)^4, \end{aligned} \quad (28)$$

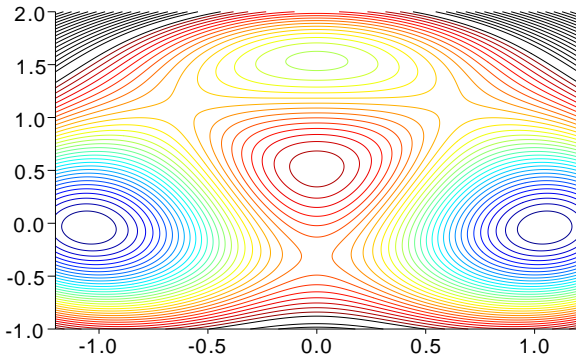
and the normalization constant

$$Z = \int_{\mathbf{X}} e^{-\beta U(x_1, x_2)} dx_1 dx_2.$$

We introduce  $d$  strata  $\mathbf{X}_\ell = (a_\ell, a_{\ell+1}) \times \mathbb{R}$ , with  $a_\ell = -R + 2(\ell - 1)R/d$  and  $\ell = 1, \dots, d$ . Thus, the element  $x = (x_1, x_2) \in \mathbf{X}$  lies in the stratum  $I(x_1) = \lfloor \frac{x_1 + R}{2R} d \rfloor + 1$ .

A plot of the level sets of the potential  $U$  is presented in Figure 1. The global minima of the potential  $U$  are located at the points  $x_- \simeq (-1.05, -0.04)$  and  $x_+ \simeq (1.05, -0.04)$  (notice that the potential is symmetric with respect to the  $y$ -axis).

The Metropolis-Hastings kernels are constructed using isotropic proposal moves distributed in each direction according to independent Gaussian random variables with variance  $\sigma^2$ . The reference Metropolis-Hastings dynamics  $P_{(\frac{1}{d}, \dots, \frac{1}{d})}$  is ergodic and reversible with respect to the measure with density  $\pi$ . This dynamics is metastable: for local moves ( $\sigma$  of the order of a fraction of  $\|x_+ - x_-\|$ ), it takes a lot of time to go from the left to the right, or equivalently from the right to the left. More precisely, there are two main metastable states: one located around  $x_-$ , and another one around  $x_+$ . These two states are separated by a region of low probability. The metastability of the dynamics increases with  $\beta$



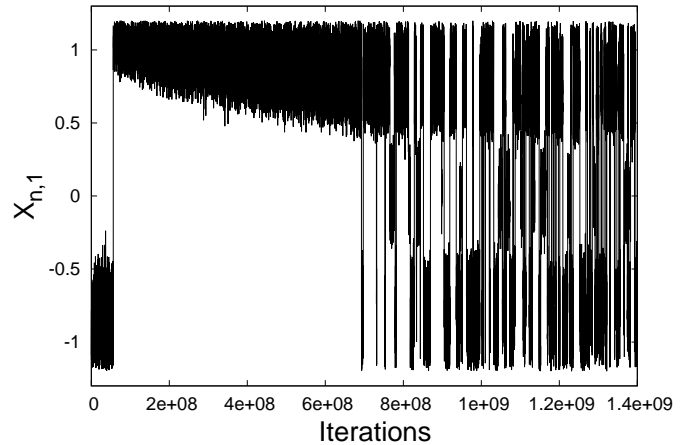
**Fig. 1** Level sets of the potential  $U$  defined in (28). The minima are located at the positions  $x_{\pm} \simeq (\pm 1.05, -0.04)$ , and there are three saddle-points, at the positions  $x_{\pm}^{\text{sd},1} \simeq (\pm 0.6, 1.15)$  and  $x^{\text{sd},2} \simeq (0, -0.3)$ . The energy differences of these saddle points with respect to the minimal potential energy are respectively  $\Delta U^1 \simeq 2.2$  and  $\Delta U^2 \simeq 2.7$ .

(i.e. as the temperature decreases). The larger  $\beta$  is, the larger is the ratio between the weight under  $\pi$  of the strata located near the main metastable states and the weight under  $\pi$  of the transition region around  $x_1 = 0$ , and the more difficult it is to leave the left metastable state to enter the one on the right (and conversely). We refer for example to [9, Fig. 3-1] for a numerical quantification of this statement.

As already pointed out in [9], adaptive algorithms such as the Wang-Landau dynamics are less metastable than the original Metropolis-Hastings dynamics, in the sense that the typical time to leave a metastable state is much smaller thanks to the adaption mechanism. In this section, we compare the adaptive Markov chain  $(X_n)_{n \geq 0}$  corresponding to the SHUS algorithm with the one generated by the Wang-Landau dynamics  $(X_n^{\text{WL}})_{n \geq 0}$  with a nonlinear update of the weights (see (11)) with stepsizes  $\gamma_n^{\text{WL}} = \gamma_*/n$  for some constant  $\gamma_* > 0$ . This choice for  $\gamma_n^{\text{WL}}$  is motivated by the asymptotic behavior of  $(\gamma_n)_{n \geq 1}$  in the limit  $n \rightarrow \infty$ , see Corollary 2. The proposal kernel used in the Metropolis algorithm is the same for SHUS and Wang-Landau. Therefore, the two algorithms only differ by the update rules of the weight sequence  $(\theta_n)_{n \geq 0}$ . The initial weight vector  $\tilde{\theta}_0$  is  $(1/d, \dots, 1/d)$  and the initial condition is  $X_0 = (-1, 0)$  for both dynamics.

#### 4.2 Study of a typical realization

Let us first consider a typical realization of the SHUS algorithm, in the case when  $\sigma$  is equal to the width of a stratum  $2R/d$ , with  $d = 48$  (so that  $\sigma = 0.05$ ),  $\gamma = 1$  and an inverse temperature  $\beta = 10$ . The values of the



**Fig. 2** Top: Typical trajectory  $X_{n,1}$  for the parameters  $d = 48$ ,  $\sigma = 2R/d = 0.05$ ,  $\gamma = 1$  and  $\beta = 10$ .

first component of the chain as a function of the iterations index  $n \mapsto X_{n,1}$  are reported in Figure 2. The trajectory is qualitatively very similar to the trajectories obtained with the Wang-Landau algorithm (see for example [9, Figure 5]). In particular, the exit time out of the second visited well is much larger than the exit time out of the first one. Such a behavior is proved for the Wang-Landau algorithm with a deterministic step-size  $\gamma_n = \gamma_*/n$  in [9, Section 6]. After the exit out of the second well, the convergence of the sequence  $(\theta_n)_{n \geq 0}$  is already almost achieved, and the dynamics freely moves between the two wells.

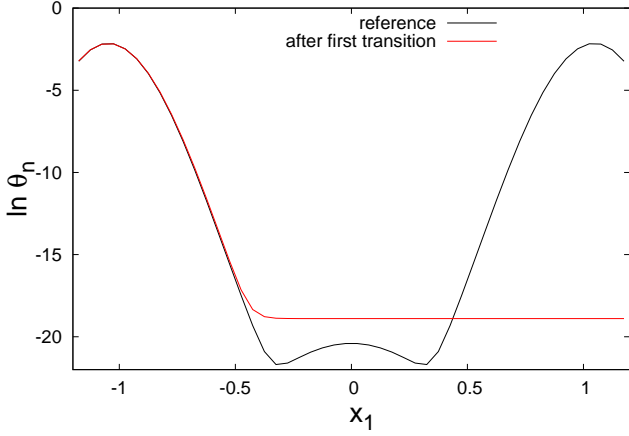
When the initial exit out of the first well occurs, the biases within this well are already very well converged, see Figure 3.

In the longtime limit, according to Figure 4 (top), one has  $n\gamma_n \rightarrow d = 48$ , as predicted by Corollary 2. In the initial phase (before the exit out of the left well),  $n\gamma_n$  stabilizes around a value corresponding to the number of strata explored within the well, see Figure 4 (bottom) (here, the exploration is well performed for values of  $x_1$  between -1.2 and -0.45, which corresponds to 15 strata). In this phase, only the restriction of the target density  $\pi$  to these strata is seen by the algorithm and this convergence can be seen as a local version of Corollary 2.

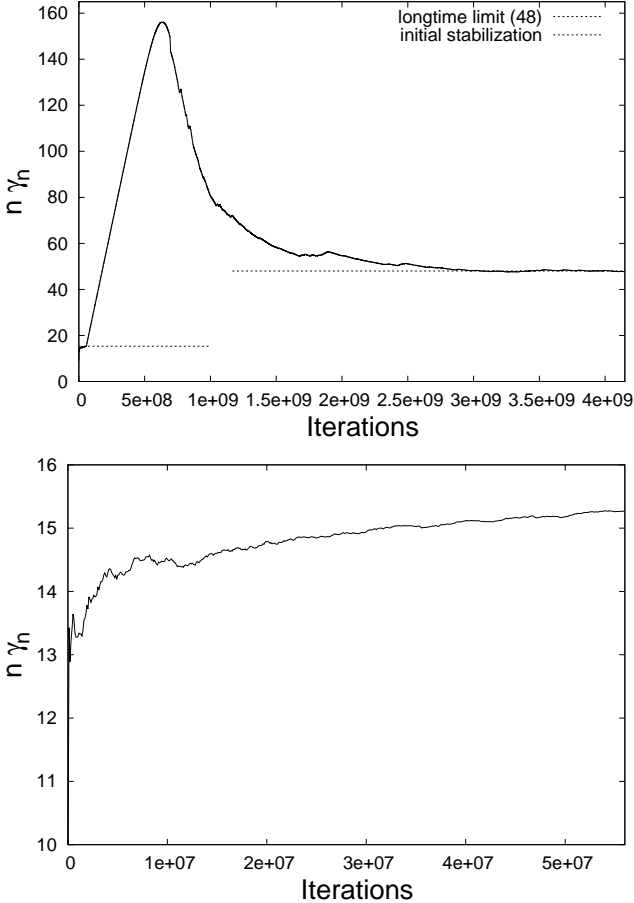
#### 4.3 Scalings of the first exit time

In this section, we study the influence of the three parameters  $\sigma$ ,  $\gamma$  and  $d$  on the first exit times (in the limit of small temperature). Concerning  $\tilde{\theta}_0$ , we stick to the assumption that without any prior knowledge on the system, the choice  $\tilde{\theta}_0(1) = \dots = \tilde{\theta}_0(d)$  is natural. In addition, notice that multiplying  $\tilde{\theta}_0$  and the parameter  $\gamma$  by





**Fig. 3** Plot of the bias  $\ln(\theta_n(I(x_1)))$  at the iteration index  $n$  when the system leaves the left well for the first time ( $d = 48$ ,  $\sigma = 2R/d = 0.05$ ,  $\gamma = 1$  and  $\beta = 10$ ). The reference values  $\ln(\theta_*(I(x_1)))$  are computed with a numerical quadrature of the integrals (1).



**Fig. 4** Behavior of the stepsize sequence  $(\gamma_n)_{n \geq 1}$  ( $d = 48$ ,  $\sigma = 2R/d = 0.05$ ,  $\gamma = 1$  and  $\beta = 10$ ). Top: Longtime convergence. Bottom: Stabilization during the exploration of the left well before the first exit.

the same constant  $c > 0$  does not modify the sequence  $(\theta_n, X_n)_{n \geq 0}$  generated by the SHUS algorithm. This is why we always choose  $\hat{\theta}_0(1) = \dots = \hat{\theta}_0(d) = 1/d$ .

Average exit times are obtained by performing independent realizations of the following procedure: initialize the system in the state  $X_0 = (-1, 0)$ , and run the dynamics until the first time index  $\mathcal{N}$  such that  $X_{\mathcal{N},1} > 1$  (*i.e.* the first component of  $X_{\mathcal{N}}$  is larger than 1) for SHUS or  $X_{\mathcal{N},1}^{\text{WL}} > 1$  for Wang-Landau. The average of this first exit time is denoted by  $t_\beta$  for SHUS and by  $t_\beta^{\text{WL}}$  for Wang-Landau. For a given value of the inverse temperature  $\beta$ , the computed average exit times  $t_\beta$  and  $t_\beta^{\text{WL}}$  are obtained by averaging over  $K$  independent realizations of the process started at  $X_0$ . We use the Mersenne-Twister random number generator as implemented in the GSL library. Since we work with a fixed maximal computational time (of about a week or two on our computing machines with our implementation of the code),  $K$  turns out to be of the order of a few thousands for the largest exit times, while  $K = 10^5$  in the easiest cases corresponding to the shortest exit times. In our numerical results, we checked that  $K$  is always sufficiently large so that the relative error on  $t_\beta$  and  $t_\beta^{\text{WL}}$  is less than a few percents in the worst cases.

According to the numerical experiments performed in [9] that confirm the theoretical analysis of a simple three-states model also given in [9], the scaling behavior for  $t_\beta^{\text{WL}}$  in the limit  $\beta \rightarrow \infty$  for the Wang-Landau algorithm with a stepsize sequence  $(\gamma_*/n)_{n \geq 1}$  is

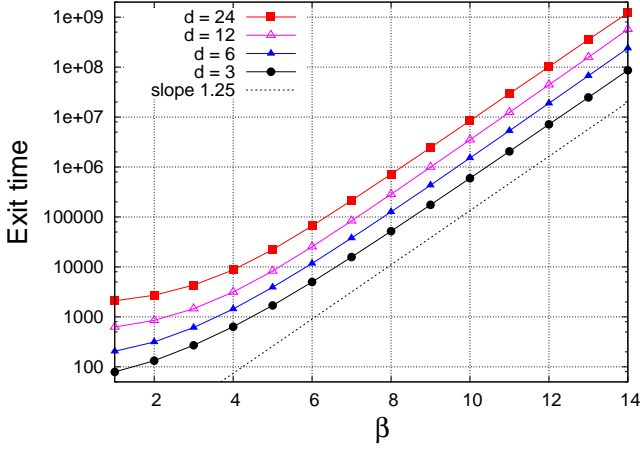
$$t_\beta^{\text{WL}} \sim C_{\text{WL}} \exp(\beta \mu_{\text{WL}}), \quad (29)$$

where  $C_{\text{WL}}$  and  $\mu_{\text{WL}}$  are positive constants which depend on  $\sigma$ ,  $\gamma_*$  and  $d$ .

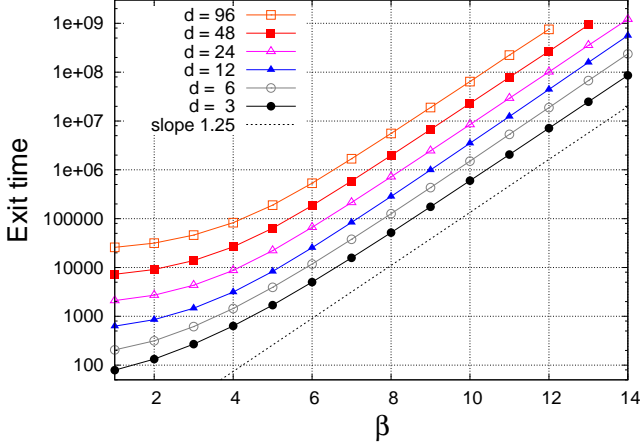
Due to the initial convergence of  $(n\gamma_n)_{n \geq 1}$  to the number  $d_{\text{sv}}$  of strata visited before the first exit time, one expects the first exit time of the SHUS algorithm to behave like the first exit time for the Wang-Landau algorithm with stepsizes  $(d_{\text{sv}}/n)_{n \geq 1}$ . Note that since we use strata of equal sizes, the number  $d_{\text{sv}}$  is proportional to the total number  $d$  of strata (see Table 4 for a more quantitative assessment).

We consider three situations:

- (i) We first study how the exit times vary as a function of  $d$  with  $\sigma = 2R/d$  and the fixed value  $\gamma = 1$ ; see Figure 5 and Table 1.
- (ii) We then fix  $\sigma = 0.1$  and study how the exit times depend on  $d$ , still with the fixed value  $\gamma = 1$ ; see Figure 6 and Table 2.
- (iii) We finally study the scaling of the exit times depending on the value  $\gamma$  when  $d = 12$  and  $\sigma = 2R/d = 0.2$  are fixed; see Figure 7 and Table 3.



**Fig. 5** Exit times as a function of the inverse temperature  $\beta$  when the number of strata  $d$  is varied while the magnitude  $\sigma$  of the proposed displacement is modified accordingly as  $\sigma = 2R/d$  (with  $\gamma = 1$  fixed).

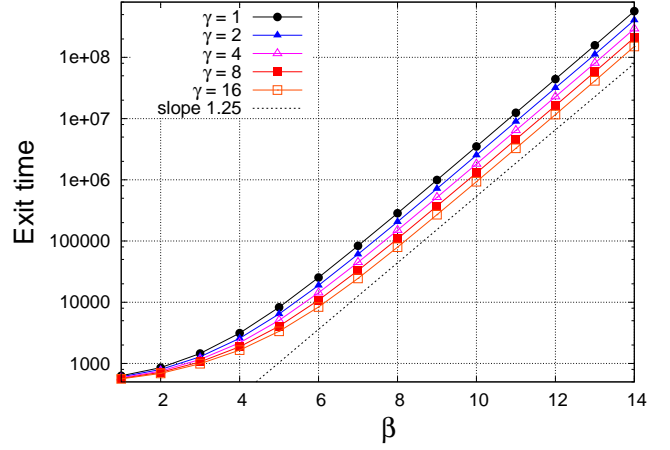


**Fig. 6** Exit times as a function of the inverse temperature  $\beta$  when the number of strata  $d$  is varied, with the fixed values  $\gamma = 1$  and  $\sigma = 0.1$ .

We observe in all cases that the average first exit time is of the form

$$t_\beta \sim C(\gamma, d, \sigma) \exp(\beta\mu)$$

in the limit of large  $\beta$ , the values  $\mu$  and  $C(\gamma, d, \sigma)$  being obtained by a least-square fit in log-log scale. In view of (29), this confirms the validity of the above comparison with the Wang-Landau algorithm. The exponential rate  $\mu$  does not seem to depend on the values of the parameters  $\sigma$ ,  $d$  and  $\gamma$ . Only the prefactors  $C(\gamma, d, \sigma)$  depend on these parameters. The larger the number of strata, and the lower the value of  $\gamma$ , the larger the prefactor is. A more quantitative assessment of the increase of the prefactor with respect to larger numbers of strata  $d$  and smaller values of  $\gamma$  is provided in the captions of Table 2 and 3.



**Fig. 7** Exit times as a function of the inverse temperature  $\beta$  when  $\gamma$  is varied, with the fixed values  $d = 12$  and  $\sigma = 2R/d = 0.2$ .

**Table 1** Scaling law  $t_\beta \sim C(\gamma, d, \sigma)e^{\beta\mu}$  for Figure 5.

	slope $\mu$	prefactor $C(\gamma, d, \sigma)$
$d = 3$	1.24	2.56
$d = 6$	1.26	5.27
$d = 12$	1.27	11.1
$d = 24$	1.24	34.9

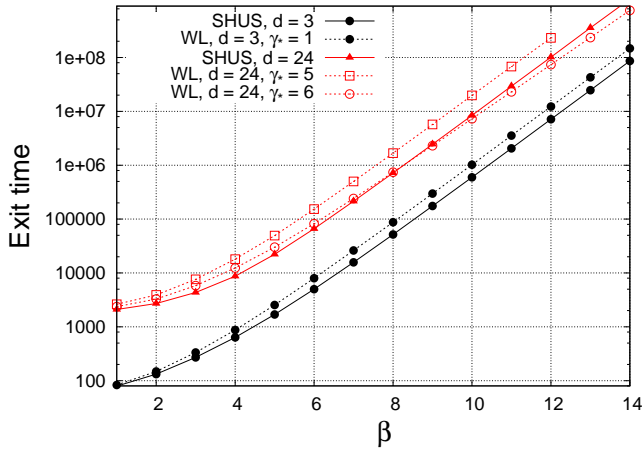
**Table 2** Scaling law  $t_\beta \sim C(\gamma, d, \sigma)e^{\beta\mu}$  for Figure 6. It appears that  $C(\gamma, d, \sigma) \simeq C(\gamma, d_0, \sigma)(d/d_0)^{1.4}$  for some reference value  $d_0$ .

	slope $\mu$	prefactor $C(\gamma, d, \sigma)$
$d = 3$	1.24	2.48
$d = 6$	1.26	5.00
$d = 12$	1.27	10.8
$d = 24$	1.24	34.9
$d = 48$	1.23	102
$d = 96$	1.23	295

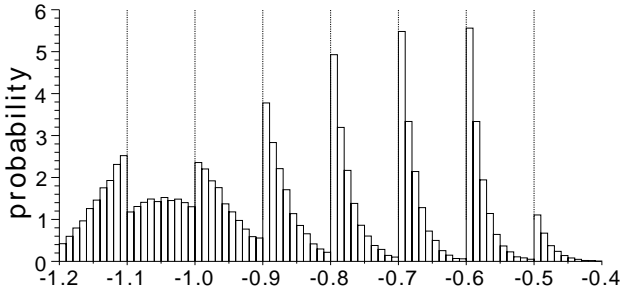
**Table 3** Scaling law  $t_\beta \sim C(\gamma, d, \sigma)e^{\beta\mu}$  for Figure 7. It appears that  $C(\gamma, d, \sigma) \simeq C(1, d, \sigma)\gamma^{-1/2}$ .

	slope $\mu$	prefactor $C(\gamma, d, \sigma)$
$\gamma = 1$	1.27	10.8
$\gamma = 2$	1.27	7.97
$\gamma = 4$	1.27	5.69
$\gamma = 8$	1.27	4.01
$\gamma = 16$	1.27	2.98

To conclude these numerical investigations on the comparison between SHUS and Wang-Landau, we look for which values of  $\gamma_*$  the two average first exit times  $t_\beta^{\text{WL}}$  and  $t_\beta$  behave similarly (in the limit of large  $\beta$ ). Some average exit times for SHUS and the Wang-Landau algorithm are presented in Figure 8. Table 4 gives intervals of values for  $\gamma_*$  for which the exponential rate



**Fig. 8** Exit times as a function of the inverse temperature  $\beta$ , for SHUS and Wang-Landau dynamics, with the choice  $\sigma = 2R/d$ . In the case  $d = 24$ , the exits times for SHUS are smaller than for Wang-Landau with parameter  $\gamma_* = 5$ , but larger than Wang-Landau with parameter  $\gamma_* = 6$ .



**Fig. 9** Histogram of the values of  $x_1$  visited along a typical trajectory before the first exit of the left metastable state, for  $\beta = 14$ ,  $d = 24$  and  $\sigma = 2R/d = 0.1$ . The frontiers of the strata are indicated by vertical lines. In this example,  $d_{sv} = 7 - 8$  since 7 strata are very well visited, while the stratum corresponding to  $-0.5 \leq x_1 \leq -0.4$  is somewhat less visited.

of increase of the exit times for the Wang-Landau dynamics matches the ones for the SHUS dynamics.

As the number of strata is increased, the value of  $\gamma_*$  has to be increased in order to retrieve the same asymptotic scaling of exit times. More precisely, we observe that  $\gamma_*$  should be proportional to  $d$  and this confirms the above comparison between SHUS and the Wang-Landau algorithm with stepsizes  $(d_{sv}/n)_{n \geq 1}$  (since  $d_{sv}$  is indeed proportional to  $d$ ). It is possible to estimate more precisely the number  $d_{sv}$  of strata visited before the first exit time by a graphical inspection, as illustrated in Figure 9. With this estimate of  $d_{sv}$ , we observe that  $\gamma_*$  is indeed very close to  $d_{sv}$ , although a bit smaller.

**Table 4** Values of  $\gamma_*$  for which the asymptotic exponential rate of increase of the exit times for Wang-Landau match the exponential rate of the SHUS dynamics, and values of  $d_{sv}$  obtained at  $\beta = 14$  by a graphical inspection (see Figure 9).

	$d_{sv}$	equivalent $\gamma_*$
$d = 3$	1	1
$d = 6$	2	1.6 – 1.8
$d = 12$	4	2.5 – 3
$d = 24$	7 – 8	5 – 6

## 5 Discussion, perspectives and extensions

### 5.1 Comparison of SHUS and the Wang-Landau algorithm

In this section, we would like to summarize our findings about the comparison between SHUS and the Wang-Landau algorithm. As explained in Section 2, SHUS can be seen as a Wang-Landau algorithm with the stepsize sequence  $(\gamma_n)_{n \geq 1}$  defined by (12). From Corollary 2, we thus expect that SHUS behaves in the longtime regime like the Wang-Landau algorithm with stepsize  $\gamma_n^{WL} = \frac{d}{n}$ . These predictions drawn from our theoretical analysis have been confirmed in the previous section by numerical experiments. Consistently, it has been shown that in terms of first exit times from a metastable state, SHUS and the Wang-Landau algorithm with  $\gamma_n^{WL} = \frac{\gamma_*}{n}$  have similar behaviors,  $\gamma_*$  being close to the number of strata visited in the metastable state containing the initial condition  $X_0$ .

We have observed numerically that the average exit time out of a metastable state for the SHUS algorithm is not drastically modified when changing the numerical parameters  $d$ ,  $\sigma$  and  $\gamma$  (see Tables 1, 2 and 3 where  $\mu$  remains approximately constant). Moreover, it is known that, in the longtime regime, the choice  $\gamma_* = d$  is the optimal one for Wang-Landau for stepsize sequences of the form  $\frac{\gamma_*}{n}$  in terms of asymptotic variance of the weight sequence, see the discussion after Theorem 3.6 in [8]. This can be seen as advantages of SHUS over Wang Landau for which, in particular, a substantial increase in the exponential rate of the exit time is observed when  $d$  is increased while  $\gamma_*$  is fixed (see [9, Table 1]).

On the downside, we observe that the scaling  $\gamma_n^{WL} \sim \frac{\gamma_*}{n}$  (when  $n \rightarrow \infty$ ) is usually not the best one in terms of efficiency. As explained in [8], convergence is also obtained for larger stepsizes  $\gamma_n^{WL} = \frac{\gamma_*}{n^\alpha}$  with  $\alpha \in (1/2, 1)$ , which allow much smaller average exit times from metastable states, see [9, Figure 3]. In this respect, SHUS is not the most efficient Wang-Landau type algorithm.

## 5.2 Accelerating SHUS

In view of the efficiency results of the Wang-Landau dynamics (see [9] for an analysis of the exit times from metastable states) and according to general prescriptions for stochastic approximation algorithms, it seems better to aim for a stepsize sequence  $(\gamma_n)_{n \geq 1}$  which decreases at the slowest possible rate while still guaranteeing convergence. In the polynomial schedule, this means that  $\gamma_n = \gamma_*/n^\alpha$  with  $\alpha$  close to  $1/2$  rather than  $\alpha = 1$ . However, it is also known that the asymptotic variance scales as  $\gamma_n$  (see for example [8, Theorem 3.6]) which calls for a very fast decaying stepsize sequence  $(\gamma_n)_{n \geq 1}$  while still guaranteeing convergence. A good practical compromise is then to combine the stochastic approximation algorithms with an averaging technique [7, 19]. This allows to use a large stepsize sequence (which yields good exploration properties) while keeping a variance of the optimal order  $1/n$ .

In view of the above discussion, a natural question is whether SHUS can be modified in order to obtain an effective stepsize sequence which scales like  $n^{-\alpha}$  with  $\alpha \in (1/2, 1)$ . A possible way of doing so is to modify the updating rule (10) as

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) \left( 1 + \frac{\gamma(\alpha) \mathbf{1}_{X_i}(X_{n+1})}{\ln \left( 1 + \sum_{j=1}^d \tilde{\theta}_n(j) \right)^{\frac{\alpha}{1-\alpha}}} \right) \quad (30)$$

for some positive deterministic  $\gamma(\alpha)$ . We call  $\text{SHUS}^\alpha$  the algorithm which consists in choosing deterministic  $X_0 \in \mathbf{X}$  and  $\tilde{\theta}_0 = (\tilde{\theta}_0(1), \dots, \tilde{\theta}_0(d)) \in (\mathbb{R}_+^*)^d$  then iterating the following steps:

**Algorithm 3** Given  $(\tilde{\theta}_n, X_n) \in (\mathbb{R}_+^*)^d \times \mathbf{X}$ ,  
 – compute the probability measure on  $\{1, \dots, d\}$ ,

$$\theta_n = \frac{\tilde{\theta}_n}{\sum_{j=1}^d \tilde{\theta}_n(j)} \in \Theta,$$

– draw  $X_{n+1}$  according to the kernel  $P_{\theta_n}(X_n, \cdot)$ ,  
 – compute, for all  $i \in \{1, \dots, d\}$ ,  $\tilde{\theta}_{n+1}(i)$  given by (30).

$\text{SHUS}^\alpha$  can be seen as a Wang-Landau algorithm with nonlinear update of the weights and with stochastic stepsizes

$$\text{for } n \in \mathbb{N}, \gamma_{n+1} = \frac{\gamma(\alpha)}{\ln \left( 1 + \sum_{j=1}^d \tilde{\theta}_n(j) \right)^{\frac{\alpha}{1-\alpha}}}. \quad (31)$$

**Proposition 4** Under A1 and A2, for each  $\alpha \in (\frac{1}{2}, 1)$ , the  $\text{SHUS}^\alpha$  Algorithm 3 converges in the sense of Definition 1. Moreover its stepsize sequence  $(\gamma_n)_{n \geq 1}$  defined by (31) satisfies

$$\mathbb{P} \left( \lim_{n \rightarrow \infty} n^\alpha \gamma_n = \gamma(\alpha)^{1-\alpha} d^\alpha (1-\alpha)^\alpha \right) = 1. \quad (32)$$

In particular, the stepsize sequence scales like  $n^{-\alpha}$  as wanted.

*Remark 3 (Relationship with the standard SHUS algorithm)* To relate the updating rule (30) with the one of SHUS (10), notice that (10) also writes

$$\begin{aligned} \tilde{\theta}_{n+1}(i) &= \tilde{\theta}_n(i) \left( 1 + \frac{\gamma \mathbf{1}_{X_i}(X_{n+1})}{\sum_{j=1}^d \tilde{\theta}_n(j)} \right) \\ &= \tilde{\theta}_n(i) \left( 1 + \frac{\gamma \mathbf{1}_{X_i}(X_{n+1})}{\lim_{\alpha \rightarrow 1^-} f \left( \alpha, \sum_{j=1}^d \tilde{\theta}_n(j) \right)} \right) \end{aligned}$$

where for  $\alpha \in (1/2, 1)$  and  $s > 0$ ,

$$f(\alpha, s) = \exp \left( \frac{\alpha}{1-\alpha} \ln \left[ 1 + (1-\alpha) \ln(1+s) \right] \right) - 1.$$

Note that, for a fixed  $\alpha \in (1/2, 1)$  and in the limit  $s \rightarrow +\infty$ ,

$$f(\alpha, s) \sim \left( (1-\alpha) \ln(1+s) \right)^{\frac{\alpha}{1-\alpha}}.$$

$\text{SHUS}^\alpha$  is therefore expected to have the same asymptotic behavior in the limit  $n \rightarrow +\infty$  as the algorithm based on the more complicated updating rule

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) \left( 1 + \frac{\gamma(\alpha)(1-\alpha)^{\frac{\alpha}{1-\alpha}} \mathbf{1}_{X_i}(X_{n+1})}{f \left( \alpha, \sum_{j=1}^d \tilde{\theta}_n(j) \right)} \right). \quad (33)$$

Notice that, with the choice,

$$\gamma(\alpha) = (1-\alpha)^{-\frac{\alpha}{1-\alpha}} \gamma, \quad (34)$$

the updating rule (33) converges to the SHUS updating rule (10) when  $\alpha \rightarrow 1^-$ . In addition, from Proposition 4,  $\lim_{n \rightarrow \infty} n^\alpha \gamma_n = \gamma^{1-\alpha} d^\alpha$  almost surely. The limiting value  $\gamma^{1-\alpha} d^\alpha$  converges to  $d$  when  $\alpha \rightarrow 1^-$  which is also consistent with what we obtained for the original SHUS algorithm, see Corollary 2.

## Numerical results

We present in this section numerical results showing that the modified SHUS algorithm (Algorithm 3) with parameter  $\alpha \in (1/2, 1)$  behaves very similarly to the Wang-Landau algorithm with stepsizes scaling as  $n^{-\alpha}$  as  $n \rightarrow +\infty$ . We choose  $\gamma(\alpha)$  according to (34) in all cases.

In order to implement the modified SHUS algorithm, some care has to be taken in order to avoid overflows related to large values of the normalization factor  $\sum_{i=1}^d \tilde{\theta}_n(i)$ . Prohibitively large numbers can be avoided

by first representing the weighted occupation factors in logarithmic scale as  $\nu_n(i) = \ln \tilde{\theta}_n(i)$ . Second, in order to avoid the uncontrolled increase of  $\nu_n(i)$  (which may be quite fast for large values of  $\beta$  and values of  $\alpha$  close to  $1/2$ ), we renormalize the factors at random stopping times where  $\sum_{i=1}^d \tilde{\theta}_n(i)$  is greater than a given (large) value  $M > 0$ . The weight sequence renormalized at these random stopping times is denoted by  $\nu_n^M(i)$ . The random stopping times are defined as  $\tau_0 = 0$ , and

$$\tau_k = \inf \left\{ n > \tau_{k-1} : \sum_{i=1}^d e^{\nu_n^M(i)} \geq M \right\} \quad (k \geq 1).$$

The renormalized factors  $\nu_n^M(i)$  evolve according to the following updating rule (obtained by taking the logarithm of (30), and possibly subtracting a renormalization factor):

$$\nu_{n+1}^M(i) = \nu_n^M(i) + \ln(1 + \gamma_{n+1}^M \mathbf{1}_{X_i}(X_{n+1})) - \sigma_n \ln M.$$

In this expression,  $\sigma_n = 0$  if  $n \neq \tau_1, \dots, \tau_k, \dots$  while  $\sigma_n = 1$  if there exists  $k \geq 1$  such that  $\tau_k = n$ . In addition, the stepsize is

$$\gamma_{n+1}^M = \frac{\gamma(\alpha)}{\left[ \ln \left( M^{-r_n} + \sum_{j=1}^d e^{\nu_n^M(j)} \right) + r_n \ln M \right]^{\frac{\alpha}{1-\alpha}}},$$

where  $r_n = \sum_{m=1}^n \sigma_m$  counts the number of times where the weights have been renormalized up to the iteration index  $n$ . The logarithmic normalized weights  $\ln \theta_n(i)$  are then constructed from the renormalized occupation measures in logarithmic scale  $\nu_n^M(i)$  as

$$\ln \theta_n(i) = \nu_n^M(i) - \ln \left( \sum_{j=1}^d e^{\nu_n^M(j)} \right).$$

The sequence of visited states  $(X_n)_{n \geq 1}$  and the sequence of weights  $(\theta_n)_{n \geq 1}$  do not depend on the value of  $M$ , as long as  $M$  is in the range of number which can be represented on a computer. In the simulations reported below, we chose  $M = 10^{10}$ . In addition, in order to have a well-behaved acceptance procedure in the Metropolis step, we compute the probability to accept the proposed move in logarithmic scale. More precisely, the proposed move  $\tilde{X}_{n+1}$  drawn from the proposal kernel starting from  $X_n$  is accepted when

$$\ln U^n \leq \ln \pi(\tilde{X}_{n+1}) - \ln \pi(X_n) - \nu_n^M(I(\tilde{X}_{n+1})) + \nu_n^M(I(X_n)),$$

where  $U^n$  is a sequence of i.i.d. uniform random variables on  $[0, 1]$ .

**Table 5** Scaling laws (35) obtained from the data presented in Figure 11.

$\alpha$	predicted value $(1 - \alpha)^{-1}$	numerical fit $\mu_\alpha$
0.6	2.50	2.47
0.7	3.33	3.33
0.8	5	5.27
0.9	10	10.8

We first check whether the effective stepsizes  $\gamma_n$  behave as predicted by Proposition 4. To this end, we simulate the modified SHUS algorithm, using increments sampled according to isotropic independent Gaussian random variables with variance  $\sigma^2$  to generate proposal moves in the Metropolis-Hastings algorithm. The value of the renormalized stepsizes

$$\frac{n^\alpha \gamma_n}{d^\alpha (1 - \alpha)^\alpha \gamma(\alpha)^{1-\alpha}} = \left( \frac{n}{d} \right)^\alpha \frac{\gamma_n}{\gamma}$$

are plotted in Figure 10. As expected, the longtime limit is 1 whatever the value of  $\alpha$ . However, the larger  $\alpha$  is, the longer it takes to attain the asymptotic regime. This is due to the fact that exit times out from the left well of the potential are typically increasing as  $\alpha$  is increased.

We study more precisely the behavior of the exit times  $t_\beta$  out from the left well in the limit of low temperatures (large  $\beta$ ), following the procedure described in Section 4.3, for various values of  $\alpha \in (1/2, 1)$ . The data presented in Figure 11 have been fitted to power laws

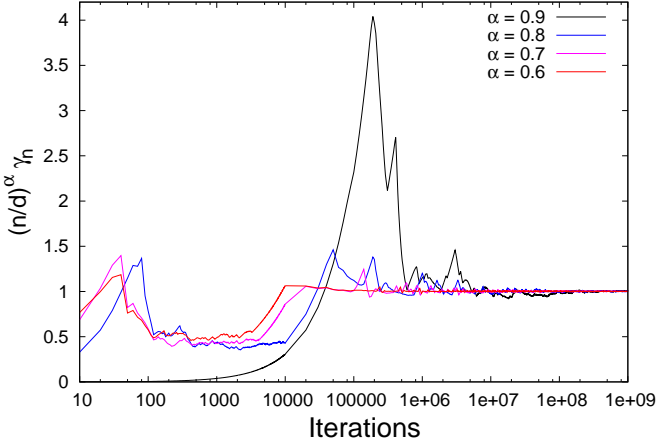
$$t_\beta \sim C_\alpha t^{\mu_\alpha}. \quad (35)$$

Such a power law scaling indeed accounts for the behavior of exit times for the standard Wang-Landau algorithm with stepsizes  $\gamma_n^{\text{WL}} = \frac{\gamma_\star}{n^\alpha}$ , and it can be proved that

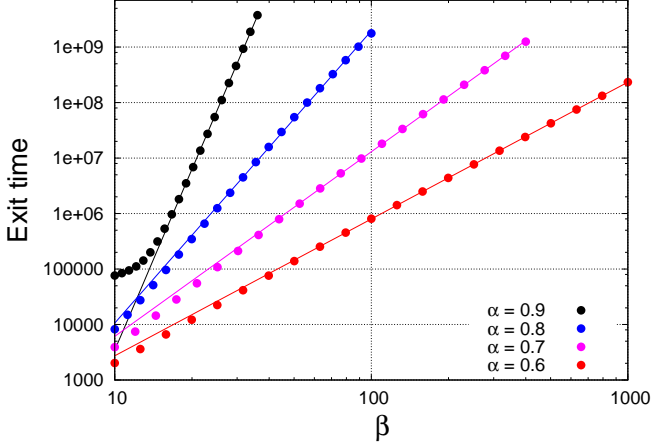
$$\mu_\alpha = \frac{1}{1 - \alpha} \quad (36)$$

for a simple model system (see [9]). This claim is also backed up in [9] by numerical experiments on the same model as the one considered in this work. The powers  $\mu_\alpha$  reported in Table 5, obtained by a least-square minimization in log-log scale, agree very well with the theoretical prediction (36). This indicates that SHUS $^\alpha$  is very close to Wang-Landau with stepsizes  $\gamma_n = \gamma/n^\alpha$ .

We next study the convergence of the logarithmic weights  $\ln \theta_n(i)$ . The empirical variance of the logarithmic weights at iteration  $n$  is estimated by running  $K$  independent realizations of the modified SHUS dynamics, as for the estimation of exit times. Denoting by  $(\ln \theta_n^k(i))_{n \geq 1}$  the weight sequence of the  $i$ th stratum for



**Fig. 10** Behavior of the sequence  $((n/d)^\alpha \gamma_n)_{n \geq 1}$  for  $d = 24$ ,  $\sigma = 2R/d = 0.1$ ,  $\gamma = 1$  and  $\beta = 10$ , for various values of  $\alpha$ . The longtime limit is in all cases 1, as predicted by Proposition 4.



**Fig. 11** Computed exit times  $t_\beta$  for the modified SHUS algorithm with various choices of the power  $\alpha$ , and the parameters  $d = 24$ ,  $\sigma = 2R/d = 0.1$ ,  $\gamma = 1$ . A power law  $t_\beta \sim C_\alpha \beta^{\mu_\alpha}$  (plotted in solid lines) is observed in all cases. Estimated powers are reported in Table 5.

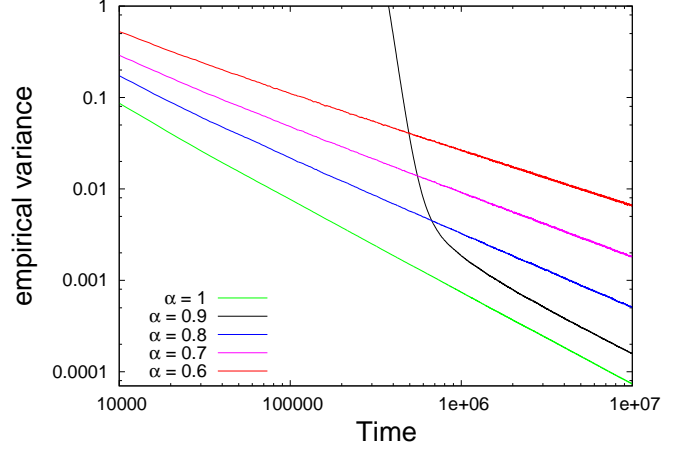
the  $k$ th realization, the empirical variance of the weight for the stratum  $i$  is estimated as

$$\mathcal{V}_{n,K}(i) = \frac{1}{K-1} \sum_{k=1}^K \left( \ln \theta_n^k(i) - \mathcal{M}_{n,K}(i) \right)^2,$$

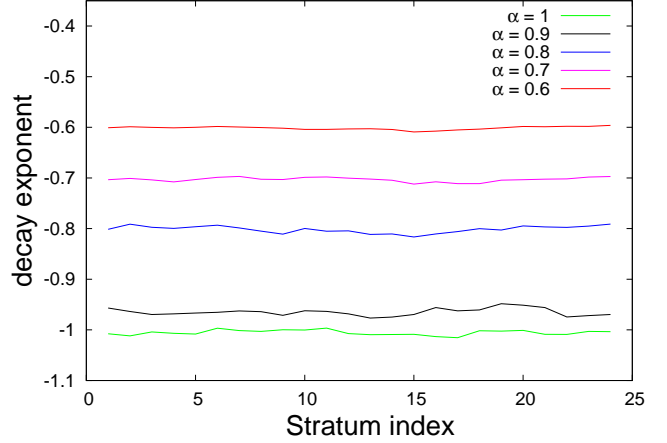
where the empirical mean  $\mathcal{M}_{n,K}(i)$  at time  $n$  is

$$\mathcal{M}_{n,K}(i) = \frac{1}{K} \sum_{k=1}^K \ln \theta_n^k(i).$$

We expect that  $\mathcal{V}_{n,K}(i)$  scales as  $\gamma_n$  in the longtime limit for  $K$  sufficiently large. Such a result is proved for the Wang-Landau dynamics in [8], where a central limit theorem is established for the weight sequence. In the sequel, we take  $K = 7 \times 10^4$ .



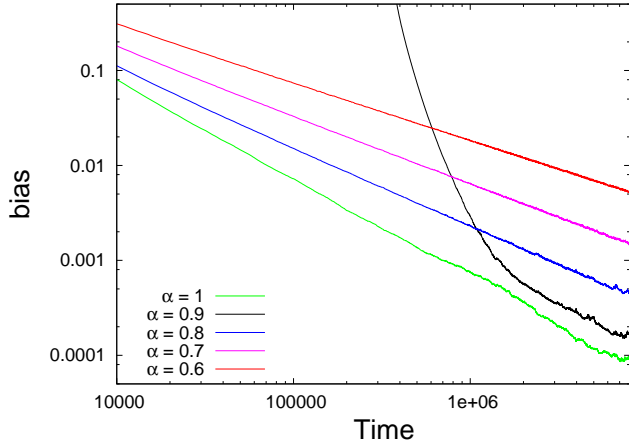
**Fig. 12** Decrease of the empirical variance  $\mathcal{V}_{n,K}(i)$  as a function of the time  $n$ , for  $d = 24$ ,  $\beta = 1$ ,  $\sigma = 2R/d = 0.1$  and  $i = 3$ . The value  $\alpha = 1$  corresponds to the standard SHUS algorithm introduced in Section 2.



**Fig. 13** Decay exponents  $a_i$  in the fit of the empirical variances  $\mathcal{V}_{n,K}(i) \sim C_{\text{var},i} n^{-a_i}$ , for various values of  $\alpha$  (same parameters as in Figure 12).

In all the computations reported below, the value  $\alpha = 1$  corresponds to the standard SHUS algorithm introduced in Section 2. The decrease of the empirical variance as a function of time is plotted in Figure 12, together with a numerical fit  $C_{\text{var},i} n^{-a_i}$  obtained by a least-square minimization in log-log scale. The decay exponents  $a_i$  for each stratum  $i$  are reported in Figure 13. We indeed confirm that the empirical variance decreases as  $n^{-\alpha}$ , except in the case  $\alpha = 0.9$  where the decrease is slightly faster than expected since the exponents  $a_i$  are around 0.95. Note also that the asymptotic regime is attained only at longer times for the value  $\alpha = 0.9$ . This is related to the fact that  $\gamma(\alpha)$  becomes very large for  $\alpha$  close to 1 (here  $\gamma(0.9) = 10^9$ ).

We finally consider the decrease of the bias in the estimated empirical averages  $\mathcal{M}_{n,K}(i)$ , as a function of



**Fig. 14** Decay of the bias  $\mathcal{B}_{n,K}$  as a function of the iteration time  $n$ , for  $d = 24$ ,  $\beta = 1$  and  $\sigma = 2R/d = 0.1$ . The corresponding decay exponents are reported in Table 6.

**Table 6** Decay of the bias  $\mathcal{B}_{n,K} \sim C_{\text{bias}} n^{-a}$  fitted on the data presented in Figure 14 for iterations times  $n$  in the range  $2 \times 10^6 \leq n \leq 8 \times 10^6$ .

$\alpha$	numerical fit $a$
0.6	0.60
0.7	0.69
0.8	0.78
0.9	0.93
1	1.01

time. We use a normalized version of the bias and average over the strata, and therefore introduce

$$\mathcal{B}_{n,K} = \sqrt{\sum_{i=1}^d \left( \frac{\mathcal{M}_{n,K}(i)}{\ln \theta_*(i)} - 1 \right)^2}.$$

The reference values  $\theta_*(i)$  are computed with a two-dimensional numerical quadrature. The decrease of the bias as a function of time is plotted in Figure 14, together with a numerical fit  $C_{\text{bias}} n^{-a}$  obtained by a least-square minimization in log-log scale. The decay exponents  $a$  are reported in Table 6. The bias approximately decrease at the same rate as the variance, namely  $n^{-\alpha}$ , a standard behavior for Monte-Carlo methods. Here again, the asymptotic behavior for the value  $\alpha = 0.9$  is observed at longer times only because of the large value of  $\gamma(\alpha)$ .

### 5.3 Partially biased dynamics

It is possible to consider more general biased measures than  $\pi_\theta$  defined in (4) by applying only a fraction of the bias. This amounts to considering the following biased

densities for a given parameter  $a \in (0, 1]$ :

$$\pi_{\theta,a}(x) = \left( \sum_{j=1}^d \frac{\theta_*(j)}{\theta(j)^a} \right)^{-1} \sum_{i=1}^d \frac{\pi(x)}{\theta(i)^a} \mathbf{1}_{\mathbf{x}_i}(x). \quad (37)$$

The motivation for applying only a fraction of the bias is to avoid having a random walk among the strata in the asymptotic regime, in order to favor the strata which are more likely under the original, unbiased measure. This idea was first proposed within the so-called well-tempered metadynamics method<sup>2</sup>, introduced in [2].

Following the reasoning of Section 2.1, Algorithm 1 should then be modified as follows:

**Algorithm 4** Given  $(\tilde{\theta}_n, X_n) \in (\mathbb{R}_+^*)^d \times \mathbf{X}$ ,

- compute the probability measure on  $\{1, \dots, d\}$ ,

$$\theta_n = \frac{\tilde{\theta}_n}{\sum_{j=1}^d \tilde{\theta}_n(j)} \in \Theta,$$

- draw  $X_{n+1}$  according to the kernel  $P_{\theta_n,a}(X_n, \cdot)$  where, for any  $\theta \in \Theta$ ,  $P_{\theta,a}$  is a transition kernel ergodic with respect to  $\pi_{\theta,a}$ ,
- compute, for all  $i \in \{1, \dots, d\}$ ,

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) + \gamma \theta_n(i)^a \mathbf{1}_{\mathbf{x}_i}(X_{n+1}). \quad (38)$$

It is possible to follow the same reasoning as in Section 2.2 to check that the only possible limit for the sequence  $(\theta_n)_{n \geq 0}$  is  $\theta_*$ . In addition,  $\gamma_n$  (defined by (12)) should scale as  $1/n$  in the longtime limit. However, the stepsize sequence  $(\gamma_n)_{n \geq 1}$  needed to rewrite the updating rule (38) as a particular case of (16) is not predictable since  $\gamma_n$  depends on  $X_n$ :

$$\gamma_n = \frac{\theta_{n-1}^a(I(X_n))}{\theta_{n-1}(I(X_n))}.$$

The convergence of this new algorithm therefore does not enter the framework of Theorem 1.

*Remark 4* The above algorithm with the modified update (38) is very much related to the well-tempered metadynamics algorithm [2]. The main difference is that we here consider a discrete reaction coordinate (associated with a partition of the state space) whereas the standard well-tempered metadynamics method is formulated for continuous reaction coordinates. As a matter of fact, in the paper [5], the authors made the observation that the well-tempered metadynamics method is

<sup>2</sup> With the notation of the metadynamics works, what we call here  $a$  is denoted  $\Delta T/(T + \Delta T)$  where  $T$  is the temperature and  $\Delta T > 0$  is a parameter. The limiting regime  $a = 1$  is recovered in the limit  $\Delta T \rightarrow +\infty$ , which corresponds to the standard metadynamics [14,3].



a stochastic approximation algorithm with a stepsize sequence of order  $1/n$  (see in particular [5, Equation (5)]).

The well-tempered metadynamics algorithm is a “parameter-free” version of the original metadynamics algorithm [14, 3]. The original metadynamics algorithm consists in penalizing already visited states in a fashion very similar to the Wang-Landau algorithm. In particular, the original metadynamics also requires to choose a vanishing stepsize sequence, in order to penalize less and less the visited states as time goes. One of the reason why the well-tempered metadynamics algorithm has then been proposed is to avoid the choice of this sequence. In the well-tempered dynamics method (as in the SHUS dynamics), the penalization decreases as the bias of the sampled point becomes larger. Roughly speaking, SHUS is a parameter-free version of Wang-Landau, in the same way as well-tempered metadynamics algorithm is a parameter-free version of metadynamics. As explained above, the parameter-free version corresponds to a specific choice of the stepsize sequence, with a  $1/n$  scaling for the strength of the penalization which does not seem to be the optimal choice in terms of exit times from metastable states, as discussed in Section 5.2.

## 6 Proofs

### 6.1 Proof of Proposition 3

For the Wang-Landau algorithm with linear update of the weights (14) and deterministic non-increasing stepsize sequence, (26) is proved in [8, Section 4.2]. It is explained in [8, Section 4.2.4] how to adapt the proof to the Wang-Landau algorithm with nonlinear update of the weights and deterministic non-increasing stepsize sequence. In addition, a careful look at the arguments in [8] shows that the randomness of the sequence  $(\gamma_n)_{n \geq 1}$  plays no role in the proof of (26) as long as the conditional distribution of  $X_{n+1}$  given  $\mathcal{F}_n$  is given by  $P_{\theta_n}(X_n, \cdot)$  and the sequence  $(\gamma_n)_{n \geq 1}$  is bounded from above by a deterministic sequence converging to 0. Hence (26) as well as the existence (proved at the end of [8, Section 4.2.1] for the Wang-Landau algorithm with linear update) of a deterministic  $p \in (0, 1)$  such that

$$\forall k, m \in \mathbb{N}, \quad \mathbb{P}(T_{k+1} - T_k > md \mid \mathcal{F}_{T_k}) \leq (1 - p)^m, \quad (39)$$

still hold in the present framework. Let us deduce from (39) that it is possible to couple a sequence  $(\tilde{T}_k)_{k \geq 0}$  with the same law as  $(T_k)_{k \geq 0}$  with a sequence  $(\tau_k)_{k \geq 1}$  of independent geometric random variables with parameter  $p$

in such a way that a.s.

$$\forall k \in \mathbb{N}, \quad \tilde{T}_{k+1} - \tilde{T}_k \leq d\tau_{k+1}. \quad (40)$$

We can set  $\tau_k = F^{-1}(U_k)$  where  $F^{-1}$  denotes the càg pseudo-inverse of the cumulative distribution function  $F(x) = \mathbf{1}_{\{x \geq 0\}} (1 - (1 - p)^{\lfloor x \rfloor})$  of the geometric law with parameter  $p$  and  $(U_k)_{k \geq 1}$  is a sequence of independent uniform random variables on  $[0, 1]$ . Now, define

$$F_{(T_0, \dots, T_k)}(x) = \mathbb{P}\left(\frac{T_{k+1} - T_k}{d} \leq x \mid (T_0, \dots, T_k)\right)$$

for  $k \geq 0$ . Since the random vector  $(T_0, \dots, T_k)$  is  $\mathcal{F}_{T_k}$ -measurable, for any  $x \geq 0$ ,

$$\begin{aligned} F_{(T_0, \dots, T_k)}(x) &\geq F_{(T_0, \dots, T_k)}(\lfloor x \rfloor) \\ &= 1 - \mathbb{E}\left[\mathbb{P}(T_{k+1} - T_k > \lfloor x \rfloor d \mid \mathcal{F}_{T_k}) \mid (T_0, \dots, T_k)\right] \\ &\geq 1 - (1 - p)^{\lfloor x \rfloor} = F(x). \end{aligned}$$

where we used (39) for the second inequality. The sequence  $(\tilde{T}_k)_{k \geq 0}$  defined inductively by  $\tilde{T}_0 = 0$  and

$$\forall k \in \mathbb{N}, \quad \tilde{T}_{k+1} = \tilde{T}_k + d F_{(\tilde{T}_0, \dots, \tilde{T}_k)}^{-1}(U_{k+1}),$$

satisfies the required properties: it has the same law as  $(T_k)_{k \geq 0}$  and it satisfies (40).

As a consequence,

$$\begin{aligned} \mathbb{P}\left(\limsup_{k \rightarrow \infty} \frac{T_k}{k} \leq \frac{d}{p}\right) &= \mathbb{P}\left(\limsup_{k \rightarrow \infty} \frac{\tilde{T}_k}{k} \leq \frac{d}{p}\right) \\ &\leq \mathbb{P}\left(\limsup_{k \rightarrow \infty} \frac{1}{k} \sum_{j=1}^k \tau_j \leq \frac{1}{p}\right) = 1, \end{aligned}$$

by the strong law of large numbers for i.i.d. random variables. One then easily concludes that (27) holds.

### 6.2 Proof of Theorem 1

The proof of Theorem 1 is performed by extending the technique of proof used in [8] for the convergence of Wang-Landau. We therefore mention only the needed extensions, the main difference with [8] being the fact that the stepsizes  $\gamma_n$  are not necessarily deterministic.

Note first that Lemma 4.6, Lemma 4.7 and Lemma 4.9 in [8] remain valid since they only depend on the expression of  $\pi_\theta$ ,  $P_\theta$ . Let us prove successively the three items in Definition 1.

(i) The proof of the first item consists in verifying the sufficient conditions given in [1, Theorems 2.2.



and 2.3] for the convergence of SA algorithms. Define the mean field function  $h : \Theta \rightarrow \mathbb{R}^d$  by

$$h(\theta) = \int_{\mathbf{X}} H(x, \theta) \pi_{\theta}(dx) = \frac{\theta_{\star} - \theta}{\sum_{i=1}^d \frac{\theta_{\star}(i)}{\theta(i)}}.$$

The function  $h$  is continuous on  $\Theta$ . By [8, Proposition 4.5], the function  $V$  defined on  $\Theta$  by

$$V(\theta) = - \sum_{i=1}^d \theta_{\star}(i) \ln \left( \frac{\theta(i)}{\theta_{\star}(i)} \right)$$

is non negative, continuously differentiable on  $\Theta$  and the level set  $\{\theta \in \Theta : V(\theta) \leq M\}$  is a compact subset of the open set  $\Theta$  for any  $M > 0$ . We also have  $\langle \nabla V(\theta), h(\theta) \rangle \leq 0$  and  $\langle \nabla V(\theta), h(\theta) \rangle = 0$  if and only if  $\theta = \theta_{\star}$ . Hence, the assumption A1 of [1] is verified with  $\mathcal{L} = \{\theta_{\star}\}$ .

Under our assumptions, the conditions on the step-size sequence  $(\gamma_n)_{n \geq 1}$  in [1, Theorems 2.2 and 2.3] hold almost-surely. To apply these theorems, it is enough to prove that

$$\limsup_k \left| \sum_{\ell \geq k}^{\ell} \gamma_{n+1} \left( H(X_{n+1}, \theta_n) - h(\theta_n) + \Lambda_{n+1} \right) \right| = 0 \quad (41)$$

with probability one, where  $\Lambda_{n+1}(i)$  is defined in (22). Indeed (26), (41) and [1, Theorems 2.2] imply that Algorithm 2 is stable in the sense that a.s., the sequence  $(\theta_n)_n$  remains in a compact subset of  $\Theta$ . Then [1, Theorems 2.3] ensures its a.s convergence to  $\theta_{\star}$ .

Let us now check (41). Since  $0 \leq \theta_n(i) \leq 1$ , it holds  $|\Lambda_{n+1}| \leq \gamma_{n+1}$ . Hence,

$$\mathbb{P} \left( \forall k, \sup_{\ell \geq k} \left| \sum_{n=k}^{\ell} \gamma_{n+1} \Lambda_{n+1} \right| \leq \sum_{n \geq k} \gamma_{n+1}^2 \right) = 1$$

so that (18) implies that  $\sup_{\ell \geq k} \left| \sum_{n=k}^{\ell} \gamma_{n+1} \Lambda_{n+1} \right|$  converges to 0 a.s. as  $k \rightarrow \infty$ .

To deal with  $H(X_{n+1}, \theta_n) - h(\theta_n)$ , for each  $\theta \in \Theta$ , we introduce the Poisson equation

$$\forall x \in \mathbf{X}, \quad g(x) - P_{\theta}g(x) = H(x, \theta) - h(\theta)$$

whose unknown is the function  $g : \mathbf{X} \rightarrow \mathbb{R}$ . Under the stated assumptions, this equation admits a solution  $\hat{H}_{\theta}(x)$  which is unique up to an additive constant and it holds (see e.g. [8, Lemma 4.9.] )

$$\sup_{\theta \in \Theta} \sup_{x \in \mathbf{X}} \left| \hat{H}_{\theta}(x) \right| < \infty. \quad (42)$$

We write

$$\begin{aligned} H(X_{n+1}, \theta_n) - h(\theta_n) &= \hat{H}_{\theta_n}(X_{n+1}) - P_{\theta_n} \hat{H}_{\theta_n}(X_{n+1}) \\ &= \mathcal{E}_{n+1} + R_{n+1}^{(1)} + R_{n+1}^{(2)}, \end{aligned}$$

with

$$\begin{aligned} \mathcal{E}_{n+1} &= \hat{H}_{\theta_n}(X_{n+1}) - P_{\theta_n} \hat{H}_{\theta_n}(X_n), \\ R_{n+1}^{(1)} &= P_{\theta_n} \hat{H}_{\theta_n}(X_n) - P_{\theta_{n+1}} \hat{H}_{\theta_{n+1}}(X_{n+1}), \\ R_{n+1}^{(2)} &= P_{\theta_{n+1}} \hat{H}_{\theta_{n+1}}(X_{n+1}) - P_{\theta_n} \hat{H}_{\theta_n}(X_{n+1}). \end{aligned}$$

Let us first check, using the  $\mathcal{F}_n$ -predictability of the sequence  $(\gamma_n)_{n \geq 1}$ , that  $M_k = \mathbf{1}_{\{k \geq 1\}} \sum_{n=1}^k \gamma_n \mathcal{E}_n$  converges a.s. as  $k \rightarrow \infty$ , which will imply that a.s.

$$\limsup_k \left| \sum_{n=k}^{\ell} \gamma_{n+1} \mathcal{E}_{n+1} \right| = 0.$$

Since  $(\mathcal{E}_n)_{n \geq 0}$  is bounded by  $\sup_{\theta \in \Theta} \sup_{x \in \mathbf{X}} |\hat{H}_{\theta}(x)|$  and  $(\gamma_n)_{n \geq 1}$  is bounded by  $\bar{\gamma}_1$ , for each  $k$ ,

$$|M_k| \leq 2k \bar{\gamma}_1 \sup_{\theta \in \Theta} \sup_{x \in \mathbf{X}} |\hat{H}_{\theta}(x)|$$

and  $M_k$  is square integrable by (42). Moreover,  $\gamma_{n+1}$  is  $\mathcal{F}_n$ -measurable and the conditional distribution of  $X_{n+1}$  given  $\mathcal{F}_n$  is  $P_{\theta_n}(X_n, \cdot)$ , so that

$$\begin{aligned} \mathbb{E}(\gamma_{n+1} \mathcal{E}_{n+1} | \mathcal{F}_n) &= \gamma_{n+1} \left[ \mathbb{E} \left( \hat{H}_{\theta_n}(X_{n+1}) | \mathcal{F}_n \right) - P_{\theta_n} \hat{H}_{\theta_n}(X_n) \right] = 0 \end{aligned}$$

In conclusion,  $(M_k)_{k \geq 1}$  is a square integrable  $\mathcal{F}_k$ -martingale. Since

$$\sum_n \mathbb{E}((M_{n+1} - M_n)^2 | \mathcal{F}_n) = \sum_n \gamma_{n+1}^2 \mathbb{E}(\mathcal{E}_{n+1}^2 | \mathcal{F}_n)$$

is smaller than  $C \sum_{n \geq 1} \gamma_n^2$  by (42) and therefore finite with probability one by (18),  $(M_k)_{k \geq 1}$  converges a.s. by [11, Theorem 2.15].

We now consider the term  $R_{n+1}^{(1)}$ . By (42) and the monotonic property of  $(\gamma_n)_{n \geq 1}$ , following the same lines as in the proof of [8, Proposition 4.10] we prove that there exists a constant  $C$  such that

$$\mathbb{P} \left( \forall k, \sup_{\ell \geq k} \left| \sum_{n=k}^{\ell} \gamma_{n+1} R_{n+1}^{(1)} \right| \leq C \gamma_{k+1} \right) = 1.$$

Therefore,  $\sup_{\ell \geq k} \left| \sum_{n=k}^{\ell} \gamma_{n+1} R_{n+1}^{(1)} \right|$  tends to zero a.s. as  $k \rightarrow \infty$ .

We now consider the term  $R_{n+1}^{(2)}$ . We have  $\hat{H}_{\theta}(x) = \sum_n P_{\theta}^n(H(\cdot, \theta) - h(\theta))(x)$ ; by [10, Lemma 4.2], there exists a constant  $C$  which does not depend on  $\theta, \theta'$  (thanks

to Proposition 2 and the upper bound  $\sup_{\theta} \sup_x |H(\theta, x)| < \infty$  such that for any  $\theta, \theta' \in \Theta$ ,

$$\sup_x \left| P_{\theta} \hat{H}_{\theta} - P_{\theta'} \hat{H}_{\theta'} \right| \leq C \left( \sup_x |H(\cdot, \theta) - H(\cdot, \theta')| + \sup_{x \in X} \|P_{\theta}(x, \cdot) - P_{\theta'}(x, \cdot)\|_{TV} + \|\pi_{\theta} d\lambda - \pi_{\theta'} d\lambda\|_{TV} \right).$$

Then, by [8, Lemmas 4.6. and 4.7], there exists a constant  $C$  such that for any  $\theta, \theta' \in \Theta$

$$\sup_x \left| P_{\theta} \hat{H}_{\theta} - P_{\theta'} \hat{H}_{\theta'} \right| \leq C \left( |\theta - \theta'| + \sum_{i=1}^d \left| 1 - \frac{\theta'(i)}{\theta(i)} \right| + \sum_{i=1}^d \left| 1 - \frac{\theta(i)}{\theta'(i)} \right| \right).$$

Since  $\sup_{\theta \in \Theta} \sup_{x \in X} |H(x, \theta)| \leq 1$  and  $\mathbb{P}(\sup_n |A_{n+1}| \leq \gamma_1) = 1$ , (20) implies that  $|\theta_{n+1} - \theta_n| \leq (1 + \gamma_1)\gamma_{n+1}$  with probability one. By (19), for any  $i \in \{1, \dots, d\}$

$$\begin{aligned} & \left| 1 - \frac{\theta_n(i)}{\theta_{n+1}(i)} \right| \vee \left| 1 - \frac{\theta_{n+1}(i)}{\theta_n(i)} \right| \\ &= \frac{|\theta_{n+1}(i) - \theta_n(i)|}{\theta_n(i) \wedge \theta_{n+1}(i)} \\ &= \frac{\gamma_{n+1} |\theta_n(I(X_{n+1})) - \mathbf{1}_{X_i}(X_{n+1})|}{1 + \gamma_{n+1} (\mathbf{1}_{X_i}(X_{n+1}) \wedge \theta_n(I(X_{n+1})))} \leq \gamma_{n+1} \quad \text{a.s.} \end{aligned} \quad (43)$$

This discussion evidences that there exists a constant  $C$  such that

$$\mathbb{P} \left( \forall k, \sup_{\ell \geq k} \left| \sum_{n=k}^{\ell} \gamma_{n+1} R_{n+1}^{(2)} \right| \leq C(1 + \gamma_1) \sum_{n \geq k} \gamma_{n+1}^2 \right) = 1.$$

By (18),  $\sup_{\ell \geq k} \left| \sum_{n=k}^{\ell} \gamma_{n+1} R_{n+1}^{(2)} \right|$  tends to zero a.s. as  $k \rightarrow \infty$ .

(ii) The proof follows the same lines as the proof of [8, Theorem 3.4] and details are omitted. The only result which has to be adapted is [8, Corollary 4.8]. Combining [8, Lemmas 4.6. and 4.7] and (43), we easily obtain the existence of a finite constant  $C$  such that almost surely, for any  $n \geq 1$

$$\begin{aligned} & \|\pi_{\theta_{n+1}} d\lambda - \pi_{\theta_n} d\lambda\|_{TV} \leq C\gamma_{n+1}, \\ & \sup_{x \in X} \|P_{\theta_n}(x, \cdot) - P_{\theta_{n+1}}(x, \cdot)\|_{TV} \leq C\gamma_{n+1}. \end{aligned}$$

(iii) The proof is very similar to the proof of [8, Theorem 3.5] and is therefore omitted.

### 6.3 Proof of Proposition 1

Throughout this proof, we denote by

$$S_n = \sum_{i=1}^d \tilde{\theta}_n(i)$$

the sum of the unnormalized weights. In view of (10)

$$S_{n+1} = S_n + \gamma \theta_n(I(X_{n+1})).$$

As  $\max_{1 \leq i \leq d} \theta_n(i) \leq 1$ , a direct induction on  $n$  yields  $S_n \leq S_0 + n\gamma$ . Since by (12)

$$\gamma_{n+1} = \frac{\gamma}{S_n}, \quad (44)$$

the lower bound on  $\gamma_{n+1}$  in Proposition 1 follows. To prove the deterministic upper-bound, we remark that

$$\begin{aligned} S_{n+1}^2 &= \left( S_n + \gamma \frac{\tilde{\theta}_n(I(X_{n+1}))}{S_n} \right)^2 \\ &\geq S_n^2 + 2\gamma \tilde{\theta}_n(I(X_{n+1})) \\ &\geq S_n^2 + 2\gamma \min_{1 \leq i \leq d} \tilde{\theta}_0(i), \end{aligned} \quad (45)$$

where we used that for each  $i \in \{1, \dots, d\}$ , the sequence  $(\tilde{\theta}_n(i))_{n \geq 0}$  is non-decreasing. By induction on  $n$ , this implies that  $S_n^2 \geq S_0^2 (1 + 2n\gamma_1 \min_{1 \leq i \leq d} \theta_0(i))$  and the deterministic upper-bound follows from (44).

To prove the stochastic upper-bound, we suppose A1 and A2. The deterministic upper-bound and Proposition 3 ensure that (27) holds. Let

$$\tilde{\underline{\theta}}_n = \min_{1 \leq i \leq d} \tilde{\theta}_n(i).$$

We have  $\tilde{\underline{\theta}}_n = S_n \underline{\theta}_n$  for each  $n \in \mathbb{N}$ . Moreover the sequence  $(\tilde{\underline{\theta}}_n)_{n \in \mathbb{N}}$  is non-decreasing and such that

$$\forall k \in \mathbb{N}^*, \tilde{\underline{\theta}}_{T_k} \geq \tilde{\underline{\theta}}_{T_{k-1}} (1 + \gamma_{T_k}), \quad (46)$$

with equality when the smallest index of stratum with smallest weight  $I_n$  is constant for  $T_{k-1} \leq n \leq T_k$ . The inequality is due to the possibility that for some  $n \in \{T_{k-1} + 1, T_{k-1} + 2, \dots, T_k - 1\}$ ,  $X_n \in \mathbf{X}_{I_{n-1}}$  and  $\exists i \in \{1, \dots, I_{n-1} - 1, I_{n-1} + 1, \dots, d\}$  such  $\tilde{\theta}_{n-1}(i) < \tilde{\theta}_{n-1}(I_{n-1}) + \gamma \theta_{n-1}(I_{n-1})$  so that  $I_n \neq I_{n-1}$ . With (27) and the lower-bound in Proposition 1 (recall that  $\gamma_1 =$

$\frac{\gamma}{S_0}$ ), one deduces that

$$\begin{aligned}\tilde{\theta}_{T_k} &\geq \tilde{\theta}_0 \prod_{j=1}^k (1 + \gamma_{T_j}) \geq \tilde{\theta}_0 \prod_{j=1}^k \left(1 + \frac{\gamma}{S_0 + \gamma(T_j - 1)}\right) \\ &\geq \tilde{\theta}_0 \prod_{j=1}^k \left(1 + \frac{\gamma}{S_0 + \gamma C_T j}\right) \\ &= \tilde{\theta}_0 \prod_{j=1}^k \frac{S_0 + \gamma + C_T \gamma j}{S_0 + C_T \gamma j} \\ &= \tilde{\theta}_0 \frac{\Gamma\left(\frac{S_0 + \gamma}{C_T \gamma} + k + 1\right)}{\Gamma\left(\frac{S_0}{C_T \gamma} + k + 1\right)} \frac{\Gamma\left(\frac{S_0}{C_T \gamma} + 1\right)}{\Gamma\left(\frac{S_0 + \gamma}{C_T \gamma} + 1\right)} \\ &\sim \tilde{\theta}_0 k^{1/C_T} \frac{\Gamma\left(\frac{S_0}{C_T \gamma} + 1\right)}{\Gamma\left(\frac{S_0 + \gamma}{C_T \gamma} + 1\right)} \text{ as } k \rightarrow +\infty.\end{aligned}$$

Hence there is a positive random variable  $C$  such that, for all  $k \in \mathbb{N}$ ,  $\tilde{\theta}_{T_k} \geq C(k+1)^{1/C_T}$ . Since by (27),  $T_{\lfloor n/C_T \rfloor} \leq n$  and the sequence  $(\tilde{\theta}_n)_{n \geq 0}$  is non-decreasing, it follows that

$$\forall n \in \mathbb{N}, \quad \tilde{\theta}_n \geq \tilde{\theta}_{T_{\lfloor n/C_T \rfloor}} \geq C \left(\frac{n}{C_T}\right)^{1/C_T} \quad \text{a.s.}$$

Since by (45),  $S_n^2 \geq S_0^2 + 2\gamma \sum_{j=0}^{n-1} \tilde{\theta}_j$ , this implies that a.s., for any  $n \geq 1$ ,

$$\begin{aligned}S_n^2 &\geq S_0^2 + 2\gamma C \sum_{j=0}^{n-1} \left(\frac{j}{C_T}\right)^{1/C_T} \\ &\geq S_0^2 + 2\gamma C C_T^{-1/C_T} \int_0^{n-1} x^{1/C_T} dx \\ &= S_0^2 + 2\gamma C C_T^{-1/C_T} \frac{C_T}{1 + C_T} (n-1)^{\frac{1+C_T}{C_T}}.\end{aligned}$$

With (44), one deduces that

$$\mathbb{P}\left(\sup_{n \in \mathbb{N}} n^{\frac{1+1/C_T}{2}} \gamma_{n+1} < \infty\right) = 1.$$

#### 6.4 Proof of Proposition 4

Let us consider the SHUS $^\alpha$  algorithm for fixed  $\alpha \in (1/2, 1)$ . For notational simplicity, we omit the dependence of  $\gamma(\alpha)$  on  $\alpha$ . The proof of Proposition 4 relies on the next lemma.

**Lemma 1** *Under Assumptions A1 and A2, the sequence  $(S_n)_{n \geq 1} = \left(\sum_{i=1}^d \tilde{\theta}_n(i)\right)_{n \geq 0}$  is increasing and bounded*

*from below by a deterministic sequence  $(\underline{s}_n)_{n \geq 0}$  going to  $+\infty$  as  $n \rightarrow \infty$ , and satisfies*

$$\begin{aligned}\mathbb{P}\left(0 < \inf_{n \geq 0} (n+1)^{\alpha-1} \ln(1 + S_n)\right. \\ \left. \leq \sup_{n \geq 0} (n+1)^{\alpha-1} \ln(1 + S_n) \leq \bar{c}\right) = 1,\end{aligned}$$

where

$$\begin{aligned}\bar{c} &= \ln(1 + S_0) \\ &\times \left(1 \vee \left(\left(1 + \frac{\gamma}{(\ln(1 + S_0))^{\frac{1}{1-\alpha}}}\right)^{\frac{1}{1-\alpha}} - 1\right)^{1-\alpha}\right).\end{aligned}$$

*Proof (of Lemma 1)* The sequence  $(S_n)_{n \geq 0}$  increases according to

$$\begin{aligned}S_{n+1} &= S_n + \frac{\gamma \tilde{\theta}_n(I(X_{n+1}))}{(\ln(1 + S_n))^{\frac{\alpha}{1-\alpha}}} \\ &= S_n \left(1 + \frac{\gamma \theta_n(I(X_{n+1}))}{(\ln(1 + S_n))^{\frac{\alpha}{1-\alpha}}}\right).\end{aligned} \tag{47}$$

Since for  $i \in \{1, \dots, d\}$ , the sequence  $(\tilde{\theta}_n(i))_{n \geq 0}$  is non-decreasing, one deduces that, for all  $n \geq 0$ ,  $S_{n+1} \geq g(S_n)$ , where

$$g(x) = x + \frac{\gamma \min_{1 \leq i \leq d} \tilde{\theta}_0(i)}{(\ln(1 + x))^{\frac{\alpha}{1-\alpha}}}$$

for  $x > 0$ . Let  $(\underline{s}_n)_{n \geq 0}$  be defined inductively by  $\underline{s}_0 = S_0$  and  $\underline{s}_{n+1} = g(\underline{s}_n)$  for all  $n \geq 0$ . This sequence is increasing and goes to  $\infty$  when  $n \rightarrow \infty$  as  $x \mapsto \gamma \left[\min_{1 \leq i \leq d} \tilde{\theta}_0(i)\right] (\ln(1 + x))^{-\frac{\alpha}{1-\alpha}}$  is locally bounded away from 0 on  $(0, +\infty)$ . Moreover  $g$  is a convex function which is decreasing on  $(0, x_g)$  and increasing on  $(x_g, +\infty)$  for some  $x_g \in (0, +\infty)$ . The minimum of the function  $g$  is therefore attained at  $x_g$ . Since

$$S_1 \geq \underline{s}_1 = g(S_0) \geq g(x_g) = x_g + \frac{\gamma \min_{1 \leq i \leq d} \tilde{\theta}_0(i)}{(\ln(1 + x_g))^{\frac{\alpha}{1-\alpha}}} > x_g,$$

it follows that  $\min(S_n, \underline{s}_n) > x_g$  for any  $n \geq 1$ , and one easily checks by induction on  $n$  that  $S_n \geq \underline{s}_n$  for all  $n \geq 1$ . Unfortunately, this lower-bound is not sharp enough to bound  $((n+1)^{\alpha-1} \ln(1 + S_n))_{n \geq 0}$  from below by a positive constant (this will be proved later on).

To prove that

$$\forall n \geq 0, \quad \ln(1 + S_n) \leq \bar{c}(n+1)^{1-\alpha}, \tag{48}$$

we remark that for all  $n \in \mathbb{N}$ , using  $S_{n+1} \geq S_n$  for the first inequality, and  $S_{n+1} \leq S_n \left(1 + \frac{\gamma}{(\ln(1 + S_n))^{\frac{\alpha}{1-\alpha}}}\right)$  for

the second one,

$$\begin{aligned} \ln(1 + S_{n+1}) &\leq \ln(1 + S_n) + \ln\left(\frac{S_{n+1}}{S_n}\right) \\ &\leq \ln(1 + S_n) + \frac{\gamma}{(\ln(1 + S_n))^{\frac{1}{1-\alpha}}}. \end{aligned} \quad (49)$$

Therefore, denoting for simplicity  $l_n = (\ln(1 + S_n))^{\frac{1}{1-\alpha}}$  and using the monotonicity of the sequence  $(l_n)_{n \geq 0}$  and the convexity of  $x \mapsto (1 + x)^{\frac{1}{1-\alpha}}$  on  $\mathbb{R}_+$  for the second inequality, we get

$$\begin{aligned} l_{n+1} &\leq l_n \left(1 + \frac{\gamma}{l_n}\right)^{\frac{1}{1-\alpha}} \\ &\leq l_n \left(1 + \frac{l_0}{l_n} \left(\left(1 + \frac{\gamma}{l_0}\right)^{\frac{1}{1-\alpha}} - 1\right)\right). \end{aligned}$$

This implies by induction on  $n$  that

$$\begin{aligned} l_n &\leq l_0 + nl_0 \left(\left(1 + \frac{\gamma}{l_0}\right)^{\frac{1}{1-\alpha}} - 1\right) \\ &\leq l_0 \left(1 \vee \left(\left(1 + \frac{\gamma}{l_0}\right)^{\frac{1}{1-\alpha}} - 1\right)\right) (n + 1). \end{aligned}$$

Raising this inequality to the power  $1 - \alpha$  leads to (48), which in turn implies the lower bound

$$\forall n \geq 1, \gamma_n \geq \gamma \bar{c}^{-\frac{\alpha}{1-\alpha}} n^{-\alpha}. \quad (50)$$

To bound  $(n + 1)^{\alpha-1} \ln(1 + S_n)$  from below, we are going to adapt the proof of Proposition 1. Since  $(\gamma_n)_{n \geq 1}$  is bounded from above by the deterministic sequence  $(\gamma(\ln(1 + \underline{s}_n))^{\frac{1}{1-\alpha}})_{n \geq 1}$  which goes to 0 as  $n \rightarrow \infty$ , Proposition 3 shows that the sequence  $(T_k)_{k \geq 0}$  defined inductively by  $T_0 = 0$  and (25) satisfies (27). Moreover, (46) still holds so that, using the concavity of the logarithm and the monotonicity of the sequence  $(\gamma_n)_{n \geq 1}$  for the second inequality then setting  $c = \frac{\gamma \ln(1 + \gamma_1)}{C_T^\alpha \gamma_1 \bar{c}^{\alpha/(1-\alpha)}}$  and using (50) and (27) for the third, one has

$$\begin{aligned} \ln \tilde{\theta}_{T_k} &\geq \ln \tilde{\theta}_0 + \sum_{j=1}^k \ln(1 + \gamma_{T_j}) \\ &\geq \ln \tilde{\theta}_0 + \frac{\ln(1 + \gamma_1)}{\gamma_1} \sum_{j=1}^k \gamma_{T_j} \\ &\geq \ln \tilde{\theta}_0 + c \sum_{j=1}^k j^{-\alpha} \\ &\geq \ln \tilde{\theta}_0 + \frac{c}{1-\alpha} ((k + 1)^{1-\alpha} - 1). \end{aligned}$$

With (27), one deduces that a.s., for all  $n \in \mathbb{N}$ ,

$$\begin{aligned} \tilde{\theta}_n &\geq \tilde{\theta}_{T_{\lfloor n/C_T \rfloor}} \\ &\geq \tilde{\theta}_0 \exp\left(\frac{c}{1-\alpha} \left(\frac{(n+1)^{1-\alpha}}{C_T^{1-\alpha}} - \frac{1}{C_T^{1-\alpha}} - 1\right)\right). \end{aligned}$$

Inserting this lower-bound together with (48) into (47) and setting  $c_0 = \frac{\gamma \tilde{\theta}_0}{\bar{c}^{\frac{\alpha}{1-\alpha}}} \exp\left(-\frac{c(1+C_T^{1-\alpha})}{(1-\alpha)C_T^{1-\alpha}}\right)$  and  $c_1 = \frac{c}{(1-\alpha)C_T^{1-\alpha}}$  one gets

$$\forall n \in \mathbb{N}, S_{n+1} \geq S_n + \frac{c_0}{(n+1)^\alpha} e^{c_1(n+1)^{1-\alpha}}.$$

Since  $x \mapsto x^{-\alpha} e^{c_1 x^{1-\alpha}}$  is increasing for  $x \geq \left(\frac{\alpha}{c_1(1-\alpha)}\right)^{\frac{1}{1-\alpha}}$ , one deduces that for all  $n \geq n_1 := \left\lceil \left(\frac{\alpha}{c_1(1-\alpha)}\right)^{\frac{1}{1-\alpha}} \right\rceil$ ,

$$\begin{aligned} S_n &\geq S_{n_1} + \int_{n_1}^n c_0 x^{-\alpha} e^{c_1 x^{1-\alpha}} dx \\ &= S_{n_1} + \frac{c_0}{c_1(1-\alpha)} \left(e^{c_1 n^{1-\alpha}} - e^{c_1 n_1^{1-\alpha}}\right) \end{aligned}$$

so that  $\limsup_{n \rightarrow \infty} (n+1)^{\alpha-1} \ln(1 + S_n) \geq c_1 > 0$ . This concludes the proof since  $(n+1)^{\alpha-1} \ln(1 + S_n) > 0$  for all  $n \geq 0$ .

*Proof (of Proposition 4)* By Lemma 1, the sequence  $(\gamma_n)_{n \geq 1}$  defined by (31) is increasing, and bounded from above by the deterministic sequence  $(\gamma(\ln(1 + \underline{s}_{n-1}))^{\frac{1}{1-\alpha}})_{n \geq 1}$  which goes to 0 as  $n \rightarrow \infty$ . Moreover,

$$\mathbb{P}\left(\gamma \bar{c}^{-\frac{\alpha}{1-\alpha}} \leq \inf_{n \geq 1} n^\alpha \gamma_n \leq \sup_{n \geq 1} n^\alpha \gamma_n < +\infty\right) = 1.$$

Since  $\sum_{n \geq 1} \frac{1}{n^\alpha} = +\infty$  and  $\sum_{n \geq 1} \frac{1}{n^{2\alpha}} < +\infty$ , one deduces that (18) holds. The convergence is therefore a consequence of Theorem 1.

To prove (32), we remark that (47) implies that

$$\ln(1 + S_{n+1}) = \ln(1 + S_n) + \frac{\gamma \theta_n(I(X_{n+1}))}{(\ln(1 + S_n))^{\frac{1}{1-\alpha}}} + R_{n+1}^{(1)},$$

where, setting  $h(x) = \ln(1 + x) - x$ ,

$$\begin{aligned} R_{n+1}^{(1)} &= h\left(\frac{\gamma S_n \theta_n(I(X_{n+1}))}{(1 + S_n)(\ln(1 + S_n))^{\frac{1}{1-\alpha}}}\right) \\ &\quad - \frac{\gamma \theta_n(I(X_{n+1}))}{(1 + S_n)(\ln(1 + S_n))^{\frac{1}{1-\alpha}}}. \end{aligned}$$

From now on,  $C$  denotes a positive random variable which may change from line to line. Lemma 1 implies that

$$\forall n \geq 0, \ln(1 + S_n) \geq C(n + 1)^{1-\alpha}.$$

Since  $0 \geq h(x) \geq -x^2/2$  for all  $x > 0$  and

$$\sup_{x>0} \frac{(\ln(1+x))^{\frac{\alpha}{1-\alpha}}}{1+x} < +\infty,$$

it follows that

$$\forall n \geq 0, \left| R_{n+1}^{(1)} \right| \leq c(\ln(1+S_n))^{-\frac{2\alpha}{1-\alpha}}$$

for some deterministic constant  $c \in (0, +\infty)$ . Writing

$$\begin{aligned} (\ln(1+S_{n+1}))^{\frac{1}{1-\alpha}} &= (\ln(1+S_n))^{\frac{1}{1-\alpha}} \\ &\times \left( 1 + \frac{\gamma \theta_n(I(X_{n+1}))}{(\ln(1+S_n))^{\frac{1}{1-\alpha}}} + \frac{R_{n+1}^{(1)}}{\ln(1+S_n)} \right)^{\frac{1}{1-\alpha}} \end{aligned}$$

and remarking that  $x \mapsto \frac{1}{x^2} \left| (1+x)^{\frac{1}{1-\alpha}} - 1 - \frac{x}{1-\alpha} \right|$  is locally bounded on  $\mathbb{R}_+$ , one deduces that

$$\begin{aligned} (\ln(1+S_{n+1}))^{\frac{1}{1-\alpha}} &= (\ln(1+S_n))^{\frac{1}{1-\alpha}} + \frac{\gamma \theta_n(I(X_{n+1}))}{1-\alpha} + R_{n+1}^{(2)}, \end{aligned}$$

where, for all  $n \geq 0$ ,  $\left| R_{n+1}^{(2)} \right| \leq c'(\ln(1+S_n))^{-\frac{\alpha}{1-\alpha}}$  for some deterministic constant  $c' \in (0, +\infty)$  depending on  $S_0$ . Lemma 1 ensures the existence of a positive random variable  $C$  such that  $\forall n \geq 0, |R_{n+1}^{(2)}| \leq C(n+1)^{-\alpha}$ . One has

$$\begin{aligned} \frac{1}{n} (\ln(1+S_n))^{\frac{1}{1-\alpha}} &= \frac{1}{n} (\ln(1+S_0))^{\frac{1}{1-\alpha}} + \frac{1}{n} \sum_{k=1}^n R_k^{(2)} \\ &+ \frac{\gamma}{(1-\alpha)n} \sum_{k=1}^n \theta_{k-1}(I(X_k)). \end{aligned}$$

The first term in the right-hand side converges a.s. to 0 as  $n \rightarrow \infty$ . So does the second since

$$\frac{1}{n} \sum_{k=1}^n k^{-\alpha} \leq \frac{1}{n} \int_0^n x^{-\alpha} dx = \frac{n^{-\alpha}}{1-\alpha}.$$

The choice  $f \equiv 1$  in (17) ensures that the third term converges a.s. to  $\frac{\gamma}{d(1-\alpha)}$ . One concludes that  $n^\alpha \gamma_{n+1} = \gamma \left( \frac{1}{n} (\ln(1+S_n))^{\frac{1}{1-\alpha}} \right)^{-\alpha}$  converges a.s. to  $\gamma \left( \frac{\gamma}{d(1-\alpha)} \right)^{-\alpha}$ .

## References

1. Andrieu, C., Moulines, E., Priouret, P.: Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44**, 283–312 (2005)
2. Barducci, A., Bussi, G., Parrinello, M.: Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100**, 020,603 (2008)
3. Bussi, G., Laio, A., Parrinello, M.: Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* **96**, 090,601 (2006)
4. Chopin, N., Lelièvre, T., Stoltz, G.: Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Stat. Comput.* **22**(4), 897–916 (2012)
5. Dama, J., Parrinello, M., Voth, G.: Well-tempered metadynamics converges asymptotically. *Phys. Rev. Lett.* **112**, 240,602(1–6) (2014)
6. Dickson, B., Legoll, F., Lelièvre, T., Stoltz, G., Fleurat-Lessard, P.: Free energy calculations: An efficient adaptive biasing potential method. *J. Phys. Chem. B* **114**, 5823–5830 (2010)
7. Fort, G.: Central limit theorems for stochastic approximation with controlled Markov chain dynamics. *ESAIM: Probability and Statistics* (2014). <http://dx.doi.org/10.1051/ps/2014013>. To appear.
8. Fort, G., Jourdain, B., Kuhn, E., Lelièvre, T., Stoltz, G.: Convergence of the Wang-Landau. *Mathematics of Computation* (2014). Accepted for publication
9. Fort, G., Jourdain, B., Kuhn, E., Lelièvre, T., Stoltz, G.: Efficiency of the Wang-Landau algorithm: A simple test case. *Appl. Math. Res. Express* **2014**, 275–311 (2014)
10. Fort, G., Moulines, E., Priouret, P.: Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.* **39**(6), 3262–3289 (2012)
11. Hall, P., Heyde, P.: *Martingale Limit Theory and its application*. Academic Press (1980)
12. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970)
13. Jacob, P.E., Ryder, R.J.: The Wang-Landau algorithm reaches the Flat Histogram criterion in finite time. *Ann. Appl. Probab.* **24**(1), 34–53 (2014)
14. Laio, A., Parrinello, M.: Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A* **99**, 12,562–12,566 (2002)
15. Lelièvre, T., Rousset, M., Stoltz, G.: Free energy computations: A mathematical perspective. Imperial College Press (2010)
16. Marsili, S., Barducci, A., Chelli, R., Procacci, P., Schettino, V.: Self-healing Umbrella Sampling: A non-equilibrium approach for quantitative free energy calculations. *J. Phys. Chem. B* **110**(29), 14,011–14,013 (2006)
17. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**(6), 1087–1091 (1953)
18. Park, S., Sener, M.K., Lu, D., Schulten, K.: Reaction paths based on mean first-passage times. *J. Chem. Phys.* **119**(3), 1313–1319 (2003)
19. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30**(4), 838–855 (1992)
20. Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**(4), 341–363 (1996)
21. Wang, F., Landau, D.: Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E* **64**, 056,101 (2001)
22. Wang, F.G., Landau, D.P.: Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**(10), 2050–2053 (2001)